Linear Models for the Analysis of Longitudinal Studies

James H. Ware

The American Statistician, Volume 39, Issue 2 (May, 1985), 95-101.

Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28198505%2939%3A2%3C95%3ALMFTAO%3E2.0.CO%3B2-L

Linear Models for the Analysis of Longitudinal Studies

JAMES H. WARE*

Longitudinal investigations play an increasingly prominent role in biomedical research. Much of the literature on specifying and fitting linear models for serial measurements uses methods based on the standard multivariate linear model. This article proposes a more flexible approach that permits specification of the expected response as an arbitrary linear function of fixed and time-varying covariates so that mean-value functions can be derived from subject matter considerations rather than methodological constraints. Three families of models for the covariance function are discussed: multivariate, autoregressive, and random effects. Illustrations demonstrate the flexibility and utility of the proposed approach to longitudinal analysis.

KEY WORDS: Repeated measures; Multivariate regression; Random effects; Autoregressive models.

1. INTRODUCTION

Longitudinal studies (defined broadly as studies in which the response of each individual is observed on two or more occasions) represent one of the principle research strategies employed in medical and social science research (Goldstein 1979; Nesselroade and Baltes 1979). Longitudinal designs are uniquely suited to the study of individual change over time, including the effects of development, aging, and other factors that affect change. Despite the importance of the longitudinal study, however, satisfactory methods for the analysis of serial measurements are not readily available. The statistical literature on the analysis of serial measurements is based on the paradigm of multivariate regression, and standard statistical software packages have the same orientation. Yet longitudinal studies typically have unbalanced designs, missing data, attrition, time-varying covariates, and other characteristics that make standard multivariate procedures inapplicable.

Recent research has focused on the development of statistical methods that not only take into account the intercorrelation of serial measurements but also accommodate the complexities of typical longitudinal data sets and permit the specification of mean-value functions determined by subject matter considerations rather than by constraints introduced by the statistical methodology. This work has been based on the notion that the analysis of serial measurements should be viewed as a univariate regression analysis of responses with correlated errors. Such a formulation suggests new and more flexible approaches to modeling and parameter estimation. This article describes these methods and their implementation. Although the approach used here could also be developed for categorical outcomes or for nonlinear modeling of continuous-outcome variables, this discussion is restricted to the problem of specifying and fitting generalized linear models for serial measurements.

A distinction is sometimes drawn between longitudinal designs (in which individual participants are followed for extended periods) and repeated measures designs (in which measurements are collected over a relatively short time period, frequently under experimental conditions). In this article, however, repeated measures (and crossover) designs are regarded as a subset of longitudinal designs, and the methods described here apply directly to data collected in the repeated measures setting.

One special type of longitudinal study not considered here can be characterized more appropriately as a follow-up study. In follow-up studies, the outcome is the time until the occurrence of a particular endpoint, such as death, disease, or remission. Although such studies involve the long-term follow-up of participants, the analysis of changes in serial measurements is not of central interest. Rather the role of serial measurements in such studies is usually limited to their use as predictors of the primary outcome. Of course, a single study can include both longitudinal and follow-up components in the sense defined here. I restrict my discussion, however, to that part of such investigations in which the serial measurements are the outcome of interest.

2. THE GOALS OF LONGITUDINAL INVESTIGATIONS

For this discussion, longitudinal studies are characterized by repeated observation of individual respondents. (The terms *respondent* and *study participant* are used because of our primary emphasis on human investigation. The methods described, however, apply quite generally to longitudinal research, and the term *experimental unit* has the same technical meaning.) The metameter for the occasions of measurement may be age, time on study, or some other natural scale; alternatively, the occasions of measurement may correspond to levels of an experimental or observational variable, possibly with no natural ordering. I will often use the generic term *time* to refer to the metameter for the occasions of measurement. In broad terms, the objectives of longitudinal research are to characterize patterns of individual response and change over time and to investigate the effects of covariates on these patterns.

Longitudinal designs are superior to cross-sectional designs in several ways. First, a longitudinal study offers investigators the opportunity for controlled and uniform measurement of exposure history and other factors related to outcome. Hence the relevant information is more reliably quantified than would be possible in a cross-sectional study

*James H. Ware is Associate Professor in the Department of Biostatistics, Harvard School of Public Health, Boston, MA 02215. This work was supported in part by National Institute of General Medical Sciences Grant GM29745 and by a cooperative agreement with the SIAM Institute for Mathematics and Society and the Environmental Protection Agency. The author is indebted to Nan Laird and Tom Louis for their many insights about longitudinal analysis.

designed to collect information retrospectively. For example, studies of the effects of low birth weight and other perinatal variables on physical development typically follow children from birth to (a) ensure that the exposure variables have been collected accurately and according to a standard protocol, (b) gather information about potentially confounding factors, and (c) document attrition in the birth cohort.

Second, longitudinal designs provide information about individual patterns of change. Such data are required for the description of development and aging, for the prediction of individual changes and for the causal interpretation of relationships between individual characteristics and patterns of change. Although cross-sectional data can be used to estimate age-related changes, these estimates are based on differences between groups rather than individual patterns of change. Such comparisons confound age and cohort effects (Fienberg and Mason 1979, Cook and Ware 1983). This is a particular problem in the study of behavioral patterns subject to temporal trends. The repeated cross-sectional design (consisting of cross-sectional samples at several time points) can separate age and cohort effects, but the estimates of age effects can be biased if the population is changing over time because of migration or attrition. In studies of cognitive development in adolescents (Baltes et al. 1979) and rate of loss of pulmonary function in adults (Glindmeyer et al. 1982), and in many other settings, cross-sectional designs have proved unsatisfactory for the study of natural history.

Finally, longitudinal designs can provide more efficient estimators of some parameters than cross-sectional designs with the same number and pattern of measurements. Fundamental to a discussion of efficiency is the distinction between *within-subject* and *between-subject* covariates. Covariates that take a single value for each respondent for the entire period of observation are called *between-subject* covariates. Sex, race, and other demographic variables that remain fixed during the investigation are between-subject covariates, as are exposure or treatment variables that do not change over time. Between-subject covariates are studied through comparisons between groups of subjects. *Within-subject* covariates are those characteristics of a respondent that vary over occasions. Thus the term *time-varying covariates* is also used. In crossover studies and in some experimental applications of the repeated measures design, treatment is a within-subject covariate. More generally, however, this category may include age, behavioral status (e.g., smoking or drinking behavior), or any other characteristic of the subjects or their environment that varies over the times of measurement, either by design or haphazardly.

Sometimes a covariate varies both within and between subjects. For example, longitudinal studies of aging frequently begin with a sample of middle-aged adults and follow the selected individuals over a period of years. In such studies, information about age-related changes is obtained both from comparisons between subjects and from comparisons between occasions for a single subject. The relative importance of these two sources of information depends on the study design. Exposure and treatment variables can also have both between-subject and within-subject variation, especially in observational studies. The distinction between these two kinds of information is important in model development for longitudinal data, as is discussed further in Section 3.

Longitudinal studies typically yield more precise estimates of the effects of within-subject covariates than would a cross-sectional design with the same number of observations but on different subjects, since these estimates are functions of differences in response at different occasions. Differences between occasions will be less variable if the responses are positively correlated, as usually is true of repeated measurements. Longitudinal designs may be less efficient, however, for estimating the effects of between-subject covariates.

3. LINEAR MODELS FOR LONGITUDINAL DATA

Consider the following generic situation. Each individual, i, $i = 1, \ldots, n$, is observed on p_i occasions. The vector of responses is denoted $\mathbf{Y}_i$. Assume that $\mathbf{Y}_i$ arises from the linear model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{e}_i, \tag{1}$$

where $\mathbf{X}_i$ is the design matrix for the ith individual and $\mathbf{e}_i$ is a vector of deviations with a multivariate normal (MVN) distribution but, for the moment, unspecified covariance structure. The design matrix can contain functions of time, however defined, and both between- and within-subject covariates. When each subject is observed at the same p times and no observations are missing, we say that the design is *balanced on time*. When $\mathbf{X}_i$ is independent of i, we say that the design is *completely balanced*. As suggested by this representation, the model $\mathbf{X}_i\boldsymbol{\beta}$ for $E(\mathbf{Y}_i)$ can be considered separately from the model for the covariance matrix $\boldsymbol{\Sigma}_i = \text{cov}(\mathbf{e}_i)$. I discuss alternative models for the mean-value and covariance functions in the next two sections.

3.1 Modeling Expected Values

In many problems, the development of a linear model for the mean-value function is the primary goal of the analysis. Initial hypotheses regarding the mean-value function derive from knowledge of the subject matter setting, and analytic methods should allow flexibility in developing the linear model through data analysis. The general representation $\mathbf{X}_i\boldsymbol{\beta}$ is familiar from ordinary linear regression and noteworthy primarily in contrast to more restricted families of models that are derived from approaches based on multivariate linear regression.

Rao (1959, 1965, 1975) considered the problem of polynomial growth curve analysis of serial measurements from a single group of subjects. When every individual is observed at the same p occasions, Rao's model can be written as $E(\mathbf{Y}_i) = \mathbf{A}\boldsymbol{\beta}$, where the columns of $\mathbf{A}$ are either powers of t (time) or the orthogonal polynomials defined by the times of observation. The rank of $\mathbf{A}$ equals the degree of the polynomial growth curve plus one. In the same balanced setting, Grizzle and Allen (1969) introduced covariates by defining the expected value as

$$E(\mathbf{Y}_i) = \mathbf{A}\boldsymbol{\beta}\mathbf{x}_i, \tag{2}$$

where $\mathbf{x}_i$ is the vector of covariate values for the ith individual. If $\mathbf{x}_i$ is $q \times 1$, then $\boldsymbol{\beta}$ is $r \times q$, where r is the number of columns in $\mathbf{A}$. Model (2) can be written in the general form of model (1) by defining $\mathbf{X}_i = \mathbf{x}_i' \otimes \mathbf{A}$, where $\otimes$ denotes the tensor product, and defining $\boldsymbol{\beta}'$ as the $1 \times rq$ vector produced by writing out the elements of the $r \times q$ matrix row by row. If we think of $\mathbf{A}$ as the matrix whose columns contain powers of t_i, this representation shows that the Grizzle and Allen model assumes that every coefficient in the polynomial model depends on each element of $\mathbf{x}_i$. More important, the representation suggests that this requirement can easily be relaxed by deleting columns containing specified products of $\mathbf{x}_i$ and powers of t from the design matrix. In short, the mean-value model can be determined directly by defining the expected value of each element Y_{ij} as the desired function of the time of observation and the covariates.

This direct approach to modeling the mean-value function has some important advantages not offered by the special structure usually assumed for the growth model. First, individuals need not be observed at the same times or on the same number of occasions. Second, time-varying covariates can be included in the model, provided that their contribution to the expected response can be written linearly. Third, covariates can modify either the expected value (level) of $\mathbf{Y}_i$ or the rate of change in $E(\mathbf{Y}_i)$; the latter arises from interaction terms involving the covariate and the appropriate power (usually the first) of t. Some variables, such as cigarette smoking or exposure to air pollution, have a natural integral over time. For cigarette smoking, the cumulative exposure is often expressed in pack-years. In such cases, the integrated exposure variable may take the place of an interaction term involving time and another covariate. Finally, other generality offered by the linear model—such as the inclusion of trigonometric functions, dependence of some (but not all) responses on particular covariates, and so on—is available to the analyst. As will be shown subsequently, this generality adds no complications in estimation and testing.

The reader should be alert to the deceptive simplicity of direct specification of a linear model for the expected response. Often this specification implicitly introduces strong assumptions. For example, consider a longitudinal study in which adult subjects 30–59 years of age are enrolled and then followed for a period of years. As previously mentioned, such a study contains both cross-sectional information about the effects of age on response (based on comparisons between subjects of different ages) and longitudinal information about the effects of age (based on comparisons between responses at different ages for a single subject). If we write $E(Y_{ij})$ as a function of age without regard for this distinction, we implicitly assume equality of the longitudinal and cross-sectional effects of aging. To maintain the distinction in the analysis, separate parameterization must be introduced for the two effects—for example, by separately modeling the effects of age on the initial observation and on changes in response over time.

Longitudinal designs in which a single cohort of subjects is observed at a common set of occasions can orthogonalize between- and within-subject information, as demonstrated in the repeated measures analysis of variance. When covariates vary both between and within subjects, the analyst intending to report separately the longitudinal and cross-sectional information is faced with the usual problems of interpretation arising from nonorthogonal designs.

3.2 Modeling the Covariance Structure

When $\boldsymbol{\Sigma}_i = \text{cov}(\mathbf{e}_i)$ are known, Aitken's generalized least squares estimator of $\boldsymbol{\beta}$ can be written as

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \mathbf{X}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{X}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{Y}_i \right). \tag{3}$$

When the covariance structure is unknown, most estimation procedures lead to estimators of the form (3) with estimates substituted for the population covariance matrixes, $\boldsymbol{\Sigma}_i$. Thus the estimation problem reduces to the problem of modeling and estimating the $\boldsymbol{\Sigma}_i$. Not surprisingly, the choice of an algorithm for estimation is influenced by the form of the underlying model.

3.2.1 General Multivariate Models. When each individual is observed at the same p occasions and there is no theoretical or empirical basis for assuming special covariance structure, one need assume only that $\text{cov}(\mathbf{e}_i) = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is an arbitrary positive definite covariance matrix. This approach can be attractive even when some observations are missing or the design is moderately unbalanced across individuals. Kleinbaum (1973) investigated multivariate methods for estimating the mean and covariance matrix in unbalanced data sets. This approach breaks down, however, when the set of observation times becomes large relative to the number of individuals: Estimation of the many parameters in the covariance matrix becomes computationally burdensome, and the resulting estimators of location parameters are inefficient when simpler covariance structures apply. Thus when the data set is highly unbalanced (i.e., individuals are observed at different sets of times) or incomplete (i.e., observations are missing), or when p is large relative to n, more parsimonious models for the covariance structure must be considered. Two natural candidates for such parsimonious models are random effects and autoregressive (AR) models.

3.2.2 Random Effects Models. Random effects models for serial measurements have a long history (Wishart 1938). Rao (1965, 1975) described a family of two-stage random effects models for serial measurements and developed estimation and testing procedures for data sets balanced on time and with no between-subject covariates. In Rao's formulation, the first stage consists of a linear model conditioned on the individual growth curve parameters, $\boldsymbol{\beta}_i$. At the second stage, the growth curve parameters are assumed to depend linearly on fixed covariates. The following expresses this idea formally:

Stage 1:

$$\mathbf{Y}_i \mid \boldsymbol{\beta}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \tag{4}$$

where $\boldsymbol{\varepsilon}_i \sim MVN(\mathbf{0}, \mathbf{R}_i)$, the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}_i$.

Stage 2:

$$\boldsymbol{\beta}_i \sim MVN(\mathbf{W}_i\boldsymbol{\alpha}, \boldsymbol{\Lambda}). \quad (5)$$

This model is unnecessarily constrained, in that linkage is introduced between the design matrixes for the location parameters and the random effects. Laird and Ware (1982) used a representation developed by Harville (1975) to avoid this restriction. If we write

Stage 1':

$$\mathbf{Y}_i \mid \mathbf{b}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad (6)$$

where $\boldsymbol{\varepsilon}_i$ is distributed as before, and

Stage 2':

$$\mathbf{b}_i \sim MVN(\mathbf{0}, \boldsymbol{\Lambda}), \quad (7)$$

then $\mathbf{X}_i$ can be freely determined as the design matrix for the expectation vector. Since the two-stage model implies that

$$\boldsymbol{\Sigma}_i = \mathbf{Z}_i\boldsymbol{\Lambda}\mathbf{Z}_i' + \mathbf{R}_i, \quad (8)$$

the number of random effects and the form of $\mathbf{Z}_i$ can be chosen to fit the observed covariance matrix. As will be discussed in Section 4, computational considerations dictate simple forms for $\mathbf{R}_i$. In practice, analysts often assume that $\mathbf{R}_i = \sigma^2\mathbf{I}$.

3.2.3 AR Models. As simple models for the covariance structure, the AR models (Box and Jenkins 1970) offer a natural alternative to random effects models. First- and second-order AR models are especially attractive for serial measurements. As will be discussed shortly, it is the residuals $\mathbf{e}_i$, rather than the observations, that satisfy an AR model. Although AR models have been the object of considerable research, the typical setting for application of time series methods is a single long time series. In the longitudinal setting, the number of realizations is large, but each observation vector is usually short. Furthermore, individual series may be both irregular and interrupted. Thus approaches to parameter estimation can be quite different in the longitudinal setting.

These three models for the covariance of serial measurements—multivariate, random effects, and AR—offer a rich set of alternatives from which to choose a covariance model for a particular application. In some problems, it may even be reasonable to consider a random effects model with errors of observation arising from an AR process. The choice among these alternatives involves both goodness of fit and ease of implementation. The next section describes approaches to model specification and parameter estimation within these families of models.

4. ESTIMATION AND HYPOTHESIS TESTING

Much of the literature on the analysis of serial measurements assumes that the data set is both completely balanced ($\mathbf{X}_i$ is independent of i) and complete (no observations are missing). In that unusual situation, closed-form maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ (the parameters of $\boldsymbol{\Sigma}_i$) are often available. Rao (1959) gave the maximum likelihood solution for a polynomial growth curve model with multivariate error structure, and Grizzle and Allen (1969) gave the solution for the multivariate regression problem with expectation vector (2). Rao (1965) gave the maximum likelihood solution for the two-stage random effects model.

In the typical application, the design is neither balanced nor complete. Thus more widely applicable approaches to parameter estimation are required. The next section describes approaches to parameter estimation for the three families of covariance structures just defined and discusses the availability of software for implementing these estimation procedures. It describes both maximum likelihood estimation and noniterative alternatives based on simple estimates of covariance parameters. This discussion assumes (except in one instance) that whenever the response $\mathbf{Y}_i$ is missing, the covariates are also missing or at least not used in the analysis. Extensions to more complicated missing data patterns have not been systematically studied.

4.1 Multivariate Models

In the multivariate model, each $\boldsymbol{\Sigma}_i$ is a submatrix defined by the relevant rows and columns of the matrix $\boldsymbol{\Sigma}$, which gives the variances and covariances for all possible pairs of measurement times. If the total number of distinct measurement times is T, the likelihood for the data can be written as a function of $\boldsymbol{\beta}$ and the $T(T + 1)/2$ parameters of $\boldsymbol{\Sigma}$. In principle, this likelihood can be maximized by such standard methods as the Newton–Raphson algorithm. In practice, no readily available software package exists for such maximum likelihood estimation, except in very simple situations. If the data consist of a single incomplete sample from an MVN distribution, the E-M algorithm (Dempster et al. 1977) can be used to estimate the mean and covariance matrix of the distribution (Beale and Little 1975). This approach can be extended to repeated measures analysis of variance with a between-subject design by using the E-M algorithm to estimate the covariance matrix for each group of subjects and then using a sweep-out procedure to estimate the pooled covariance matrix (Berk 1984). In these two situations, the E-M algorithm can be implemented using BMDPAN (Dixon 1983).

When the design matrix $\mathbf{X}_i$ includes values of time-varying covariates, the algorithmic approach to maximum likelihood estimation depends on the availability of the values of the time-varying covariates when the response is missing. If the values of $\mathbf{X}_i$ are available, maximum likelihood estimates can again be obtained by a standard application of the E-M algorithm. When they are not available, as would be the case when an entire observation is missing, the solution is less straightforward. One approach combines the E-M algorithm with a modified Gauss–Seidel procedure (Dahlquist and Bjork 1974). Since the gradient equations for $\boldsymbol{\beta}$ are satisfied by (3) with $\boldsymbol{\Sigma}_i$ replaced by its maximum likelihood estimate, the pth step of the algorithm utilizes the current estimate $\hat{\boldsymbol{\beta}}^{(p)}$ of the regression coefficients to compute residuals: $\mathbf{e}_i^{(p)} = \mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{(p)}$. By treating these residuals as an incomplete sample from the MVN distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ and applying the E-M algorithm as described by Beale and Little (1975), one can obtain solutions to the gradient equations for the elements of $\boldsymbol{\Sigma}$ conditional on the current estimate of the location parameters $\hat{\boldsymbol{\beta}}^{(p)}$. Iteration of this procedure alternating between reestimation of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ produces the maximum likelihood estimates.

Since this procedure requires two levels of iteration, it is computationally intensive and therefore unsuitable for standard practice. Fortunately, $\hat{\boldsymbol{\beta}}$ as defined in (3) is often relatively insensitive to the estimate of $\boldsymbol{\Sigma}$. Thus noniterative estimates of $\boldsymbol{\Sigma}$ are a reasonable alternative to maximum likelihood. When the data set is nearly complete, $\boldsymbol{\Sigma}$ can be estimated from the complete cases. When the data are more highly unbalanced, the following three-step estimation procedure can be employed: (a) Estimate the regression parameters $\boldsymbol{\beta}$ by ordinary least squares (OLS); that is, ignore the covariance between the repeated observations of a single participant. (b) Using the residuals from (a), estimate each element of $\boldsymbol{\Sigma}$ using all available observations or pairs of observations available at the relevant times. (c) Reestimate $\boldsymbol{\beta}$ by using expression (3) with the estimates of $\boldsymbol{\Sigma}_i$ chosen appropriately from $\boldsymbol{\Sigma}$.

When the estimate of $\boldsymbol{\Sigma}$ is not positive definite (a possibility with this procedure) a positive definite matrix can be obtained by replacing negative eigenvalues of $\boldsymbol{\Sigma}$ with small positive values (Bock and Peterson 1975) or by performing one iteration of the E-M algorithm with the residuals from the OLS fit. Since even small positive eigenvalues can lead to unstable estimates of regression coefficients, a lower limit on the eigenvalues can be specified or the covariance matrix can be smoothed by using techniques analogous to ridge regression (Dixon 1983, p. 221).

4.2 Random Effects Models

Maximum likelihood (ML) estimates of the mean and covariance parameters of the random effects model can be obtained either by the E-M algorithm or by such second-order algorithms as Newton–Raphson. Laird and Ware (1982) described how the E-M algorithm is applied in this setting to obtain either ML or restricted ML (REML) estimates of location and scale parameters. Software has been developed using that approach (Cook 1982). The use of Newton–Raphson and related procedures to estimate parameters of random effects models was discussed by Jennrich and Sampson (1976). These authors, however, oriented their work toward applications in the analysis of variance. Thus their computer program BMDP3V (Dixon 1983) requires independence of different random effects and does not exploit the block diagonal covariance matrix characteristic of the longitudinal setting.

As with the multivariate model, estimation of the parameters of random effects models by maximum likelihood can be prohibitively computer intensive, especially when the data set is large. Thus approximations based on noniterative estimates of covariance parameters should be considered. For the analysis of variance setting, noniterative estimates of variance components have been extensively studied, although most of this work assumes independence of different random effects (Searle 1971). For models in which the $\mathbf{Z}_i$ depend on covariates, such as height or age, that vary both over time and between subjects (as might be appropriate for a growth curve model) no general methodology has been developed.

When the random effects arise from a two-stage growth curve model [as defined in (4) and (5)], a two-step analysis can be used to reduce computation. At the first step, the growth curve (4) is fitted for each individual, using that individual's serial observations. At the second step, the estimated growth curve coefficients are individually analyzed by weighted or unweighted least squares. If $\mathbf{R}_i = \sigma^2\mathbf{I}$, the two-stage growth curve model implies that $\operatorname{var}(\hat{\boldsymbol{\beta}}_{ij}) = \lambda_{jj} + \sigma^2(\mathbf{Z}_i'\mathbf{Z}_i)_{jj}^{-1}$. In some simple situations, λ_{jj} and σ^2 can be estimated by the method of moments, and these estimates can then be used in weighted least squares regression analysis of the $\hat{\boldsymbol{\beta}}_{ij}$. In more complicated situations, it may be reasonable to assume that either the within- or between-individual component of variance is negligible. Hui and Berger (1983) described an iterative procedure for empirical Bayes analysis of the $\hat{\boldsymbol{\beta}}_{ij}$. Fay and Herriot (1979) discussed methods for estimating the variance components in simple situations.

4.3 AR Models

The general approach to parameter estimation for AR models can be illustrated by a discussion of the first-order AR model, [AR(1)]. If $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{e}_i$, $i = 1, \ldots, n$, and $\mathbf{e}_i$ is AR(1), then the likelihood can be written in terms of Y_{i1} and Y_{ij} ($j = 2, \ldots, p_i$, $i = 1, \ldots, n$) by using the representations

$$Y_{i1} = \mathbf{X}_{i1}'\boldsymbol{\beta} + e_{i1} \tag{9}$$

and

$$Y_{ij} - \rho Y_{i,j-1} = (\mathbf{X}_{ij}' - \rho\mathbf{X}_{i,j-1}')\boldsymbol{\beta} + v_{ij}, \tag{10}$$

where $\mathbf{X}_{ij}'$ is the jth row of $\mathbf{X}_i$, ρ is the correlation between successive observations, and $v_{ij} = e_{ij} - \rho e_{i,j-1}$ is distributed as $N(0, \sigma^2(1 - \rho^2))$.

Maximum likelihood estimates of the parameters of this model can be obtained by a modified Gauss–Seidel procedure (Louis and Spiro, 1984). One part of the procedure consists of weighted least squares estimation of $\boldsymbol{\beta}$ from (9) and (10), with ρ set to its current estimate and treated as known. The other part consists of maximization of the likelihood of the residuals $Y_{ij} - \mathbf{X}_{ij}'\hat{\boldsymbol{\beta}}$ with respect to ρ, with $\hat{\boldsymbol{\beta}}$ treated as known. This is equivalent to the procedure described earlier for the multivariate setting, except that a parsimonious model is assumed for the covariance structure.

Noniterative estimates of the mean and variance components of this model can be computed by using an adaptation of the three-step procedure described previously for multivariate models. The first step proceeds as before. At the second step, σ^2 and ρ can be estimated from the sums of squares and sums of cross products, respectively, of successive residuals (Parks 1967; LaVange and Helms 1983). The regression coefficients $\boldsymbol{\beta}$ are then reestimated by using (3) with the estimates of $\boldsymbol{\Sigma}_i$ based on the assumed AR error structure and the estimated variance components. The three-step estimation procedure can be iterated by using the regression coefficients defined at the previous step to reestimate the residuals and thereby the variance components. The extension to pth order AR models is straightforward, involving a representation analogous to (9) and (10) that includes one unconditional and p conditional expressions.

This analysis assumes an AR structure for the e_{ij}, not the Y_{ij}. Although some investigators have studied models of the

form

$$Y_{ij} = \mathbf{X}'_{ij}\boldsymbol{\beta} + \rho Y_{i,j-1} + v_{ij}, \qquad (11)$$

these models are fundamentally different from (4). The regression coefficients $\boldsymbol{\beta}$ in (9) and (10) are the coefficients of the unconditional regression model (1), whereas the regression coefficients in (11) are interpreted conditionally.

When the analyst is reluctant to specify the exact form of the covariance between successive observations, moving average error structures offer an alternative to the AR model. LaVange (1983) described a three-step procedure for regression analysis of serial measurements with moving average error structure that is based on an adaptation of a procedure developed by Seely (1969).

5. CHOOSING A MODEL FOR THE COVARIANCE STRUCTURE

Given the several families of models available for the analysis of serial measurements, which model and estimation procedure should be chosen for a particular problem? Although no simple rule can be given, several of the important considerations can be identified. The first of these is feasibility. When the covariance of successive measurements depends on the times of measurement and the number of measurement times is large, the general multivariate model will either be inestimable or very inefficient. Time series models can, in principle, be fitted in this setting. For example, the first-order AR model could be taken to imply that the correlation between successive observations is ρ to the power Δt, where Δt is the difference in the times of measurement. In practice, however, software limitations make time series models difficult to fit in this setting. Random effects models can be defined in terms of polynomial functions of time, so an individual's covariance matrix depends on the pattern of measurements. Thus unbalanced data sets introduce no special difficulties for random effects models.

Interestingly, the degree of imbalance in a data set can be influenced by the choice of the metameter for time. For example, in a study of development of pulmonary function during childhood, a model assuming that covariances depend on the heights at the successive times of measurement (as might be suggested by the strong dependence of pulmonary function on height; see Dockery et al. 1983) would result in a highly unbalanced design. If covariances were assumed to depend on age, a study with regularly scheduled visits could produce a relatively balanced data set.

The second consideration is goodness of fit. Both the random effects and AR models introduce strong assumptions about covariance patterns, especially when the number of observations on each subject is large. Goodness of fit can be assessed empirically by comparing the sample covariance matrix with the fitted covariance matrix in groups of subjects with a common observation pattern. In some situations, likelihood ratio tests can be used to compare nested models for covariance structure (Rao 1965). Although random effects and time series models can be used even when the fit is poor, estimates of location parameters will be inefficient, and the estimators of their standard errors will be biased.

When a family of models has been chosen, the choice between noniterative and iterative ML estimators must be considered. In many applications, estimates of location parameters and their estimated standard errors are insensitive to refinements of estimates of scale parameters. In this situation, iterative estimation of variance components is of little value in sharpening inferences about parameters of the mean-value function. This is especially true in large data sets, where the cost of ML estimation of variance components can be substantial.

6. RESIDUAL ANALYSIS AND REGRESSION DIAGNOSTICS

Relatively little attention has been given to the development of regression diagnostics for serial measurements. For any of the models described here, standardized cross-sectional residuals can be defined in the usual way. Similarly, longitudinal residuals can be defined as deviations of successive observations from their expected values, given previous responses and the values of covariates included in the linear model. For the AR(1) model, longitudinal residuals have a particularly simple form, namely $Y_{ij} - \rho Y_{i,j-1} - (\mathbf{X}'_{ij} - \rho\mathbf{X}'_{i,j-1})\boldsymbol{\beta}$. Since this expression is of the same form for $j = 2, \ldots, p_i$, baseline residuals and follow-up residuals can be analyzed separately—for instance, by plotting the residuals against the corresponding residuals of a proposed additional covariate (Louis and Spiro 1984). Although analogous longitudinal residuals can be described for the multivariate and random effects models, their form varies from occasion to occasion and possibly from subject to subject, making residual analyses more difficult to generate and interpret. Additional work is needed to develop and implement methods for such residual analysis and for the detection of influential and outlying points.

Since the methods described here assume normality, there is also a need to assess the adequacy of that assumption. For the multivariate model, this can be approached by a variety of methods proposed for assessing the normality of a multivariate distribution, none of them wholly satisfactory. For random effects models, Ryan and Dempster (1984) proposed the weighted normal plot as a graphical approach to the assessment of normality. For simple AR models, univariate methods for the assessment of normality can be applied to the transformed variables defined by (9) and (10).

Failure of the normality assumption does not invalidate the estimates of location parameters, since weighted least squares estimates will be unbiased and consistent under very broad conditions. It does, however, invalidate the usual tests and confidence intervals based on normal theory. Since nonnormality is a common situation in applications, an approach to inference that does not rely on normality would be useful. Although some progress has been made in developing nonparametric procedures for the analysis of serial measurements (Zerbe 1979; Puri and Sen 1971), these methods are applicable only to relatively simple and balanced situations.

The bootstrap (Efron and Gong 1983) provides an alternative approach to assessing the distributional properties of estimators without reliance on normality assumptions. The bootstrap provides an estimate of the sampling distribution of an estimator based only on that estimation procedure and

the sample in hand, without appealing to distributional assumptions. If an approximately optimal estimator can be defined, the bootstrap can be used to assess its properties. The major limitation of the bootstrap in this application may prove to be its heavy use of computer time. In developing bootstrap techniques for the analysis of serial measurements, special attention should be given to the impact of missing data on inference.

[*Received May 1984. Revised October 1984.*]

REFERENCES

Baltes, P. B., Cornelius, S. W., and Nesselroade, J. R. (1979), "Cohort Effects in Developmental Psychology," in *Longitudinal Research in the Study of Behavior and Development*, eds. J. R. Nesselroade and P. B. Baltes, New York: Academic Press, pp. 61–87.

Beale, E. M. L., and Little , R. J. A. (1975), "Missing Values in Multivariate Analysis," *Journal of the Royal Statistical Society*, Ser. B, 37, 129–145.

Berk, K. (1984), "Computing for Unbalanced Repeated Measures Experiments," in *Proceedings of the Eleventh Annual SUGI Conference*, Cary, NC: SAS Institute.

Bock, R. D., and Peterson, A. C. (1975), "A Multivariate Correction for Attenuation," *Biometrika*, 62, 673–678.

Box, G. E.P., and Jenkins, G. M. (1970), *Time Series Analysis, Forecasting and Control*, San Francisco: Holden-Day.

Cook, N. R. (1982), "A General Linear Model Approach to Longitudinal Data Analysis," unpublished Ph.D. dissertation, Harvard School of Public Health, Dept. of Biostatistics.

Cook, N. R., and Ware, J. H. (1983), "Design and Analysis Methods for Longitudinal Research," *Annual Review of Public Health*, 4, 1–24.

Dahlquist, G., and Bjork, A. (1974), *Numerical Methods*, New York: Prentice-Hall.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the E-M Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Dockery, D., Berkey, C., Ware, J. H., Speizer, F. E., and Ferris, B. G., Jr. (1983), "Distribution of Forced Vital Capacity and Forced Expiratory Volume in One Second in Children Six to Eleven Years of Age," *American Review of Respiratory Diseases*, 129, 366–374.

Dixon, W. J. (ed.) (1983), *BMDP Statistical Software*, Berkeley: University of California Press.

Efron, B., and Gong, G. A. (1983), "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *The American Statistician*, 37, 170–174.

Fay, R. E., and Herriot, R. A. (1979), "Estimates of Income for Small Places," *Journal of the American Statistical Association*, 74, 269–277.

Fienberg, S. E., and Mason, W. M. (1979), "Identification and Estimation of Age-Period-Cohort Models in the Analysis of Discrete Archival Data," in *Sociological Methodology*, ed. D. R. Heise, New York: Jossey-Bass, pp. 1–67.

Glindmeyer, H. W., Diem, J. D., Jones, R. N., and Weill, H. (1982), "Noncomparability of Longitudinally and Cross-Sectionally Determined Annual Change in Spirometry," *American Review of Respiratory Diseases*, 125, 544–548.

Goldstein, H. (1979), *The Design and Analysis of Longitudinal Studies*, New York: Academic Press.

Grizzle, J., and Allen, D. (1969), "Analysis of Growth and Dose Response Curves," *Biometrics*, 25, 357–381.

Harville, D. A. (1975), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association*, 72, 320–340.

Hui, S. L., and Berger, J. O. (1983), "Empirical Bayes Estimation of Rates in Longitudinal Studies," *Journal of the American Statistical Association*, 78, 753–760.

Jennrich, R. I., and Sampson, P. F. (1976), "Newton–Raphson and Related Algorithms for Maximum Likelihood Variance Component Estimation," *Technometrics*, 10, 63–67.

Kleinbaum, D. G. (1973), "A Generalization of the Growth Curve Model Which Allows Missing Data," *Journal of Multivariate Analysis*, 3, 117–124.

Laird, N. M., and Ware, J. H. (1982), "Random Effects Models for Longitudinal Data," *Biometrics*, 38, 963–974.

LaVange, L. M. (1983), "Analysis of Incomplete Longitudinal Data With Constrained Covariance Structures," unpublished Ph.D. dissertation, University of North Carolina at Chapel Hill, Dept. of Biostatistics.

LaVange, L. M., and Helms, R. W. (1983), "The Analysis of Incomplete Longitudinal Data With Time Series Covariance Structures," paper presented at the Joint Statistical Meetings, Toronto, Canada.

Louis, T. A., and Spiro, A. III (1984), "Fitting First-Order Autoregressive Models With Covariates," manuscript submitted for publication.

Nesselroade, J. R., and Baltes, P. B. (eds.) (1979), *Longitudinal Research in the Study of Behavior and Development*, New York: Academic Press.

Parks, R. (1967), "Efficient Estimation of a System of Regression Equations When Disturbances Are Both Serially and Contemporaneously Correlated," *Journal of the American Statistical Association*, 62, 500–501.

Puri, M. L., and Sen, P. K. (1971), *Nonparametric Methods in Multivariate Analysis*, New York: John Wiley.

Rao, C. R. (1959), "Some Problems Involving Linear Hypotheses in Multivariate Analysis," *Biometrika*, 46, 49–58.

——— (1965), "The Theory of Least Squares When the Parameters Are Stochastic and Its Application to the Analysis of Growth Curves," *Biometrika*, 52, 447–458.

——— (1975), "Simultaneous Estimation of Parameters in Different Linear Models and Applications to Biometric Problems," *Biometrics*, 31, 545–554.

Ryan, L. M., and Dempster, A. P. (1984), "Weighted Normal Plots," Technical Report 394Z, Dana–Farber Cancer Institute, Boston, MA.

Searle, S. R. (1971), "Topics in Variance Component Estimation," *Biometrics*, 27, 1–76.

Seely, J. (1969), "Estimation in Finite-Dimensional Vector Spaces With Application to the Mixed Linear Model," unpublished Ph.D. dissertation, Iowa State University, Dept. of Statistics.

Wishart, J. (1938), "Growth-Rate Determinations in Nutrition Studies with the Bacon Pig, and Their Analysis," *Biometrics*, 30, 16–28.

Zerbe, G. (1979), "Randomization Analysis of the Completely Randomized Design Extended to Growth and Response Curves," *Journal of the American Statistical Association*, 74, 215–221.