Ann. Rev. Public Health. 1983. 4:1–23

DESIGN AND ANALYSIS METHODS FOR LONGITUDINAL RESEARCH

Nancy R. Cook

Harvard Medical School, Brigham and Women's Hospital, Boston, Massachusetts 02146

James H. Ware

Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115

INTRODUCTION

We describe in this article methods for the design, conduct, and analysis of longitudinal studies. We define a longitudinal study as one in which each individual is observed on more than one occasion. The observation may be a measurement, such as diastolic blood pressure, or a state, such as presence or absence of asthma symptoms. We distinguish longitudinal studies from follow-up studies, in which individuals are followed until the occurrence of an event such as death or myocardial infarction. Issues of design and analysis that are unique to follow-up studies are not discussed in this paper.

Longitudinal designs have two principal motivations.

1. To increase the precision of treatment contrasts by eliminating interindividual variation: This is achieved by observing each subject under the several treatment (or exposure) conditions to be compared. Such designs are called *repeated measures* designs, and include the *cross-over* design as a special case. Repeated measures designs use each subject as his or her own control.
2. To examine the individual's changing response over time: Longitudinal designs have natural appeal for the study of changes associated with

0163-7525/83/0510-0001-$02.00

development or aging. They have value for describing both temporal changes and their dependence on individual characteristics.

We begin by comparing the advantages of cross-sectional designs, *pure longitudinal designs* in which a single cohort is followed over time, and *mixed longitudinal designs,* in which several cohorts are followed for a shorter period. We then discuss the design of longitudinal studies, particularly the duration and frequency of measurement in studies of growth and aging, and describe methods for reducing the variability of observations. Because the data available for analysis depend on both the design and conduct of a study, we also discuss some of the major issues to be considered in the conduct of a longitudinal study.

Finally, we consider the analysis of longitudinal studies. A considerable literature exists on the analysis of serial measurements using the general linear model. We describe these methods, particularly growth curve analysis, tracking, and repeated measures analysis, and their extension to the typical situation in which the data are incomplete. We also discuss nonlinear growth curve analysis and the analysis of serial binary responses. This article is written for the health scientist who is not a statistician but has some familiarity with the Analysis of Variance (ANOVA). Although we avoid detailed technical discussions, complete and current references to the statistical literature are provided.

DESIGN

Comparing Longitudinal and Cross-sectional Designs

Longitudinal studies are usually motivated either by the desire for precise comparisons of treatments or by intrinsic interest in age- or time-related changes. But cross-sectional designs can also be used to compare treatments and to investigate age-related changes. Treatments can be administered simultaneously to separate groups of subjects. Similarly, age-specific standards for height and weight in children were originally developed from data gathered cross-sectionally. Frequently, investigators choose between alternative designs by comparing their precision, potential for bias, and feasibility. In this section, we compare the precision of longitudinal and cross-sectional designs, discuss the potential for bias due to age, time, and cohort effects, and conclude with a more general discussion of the trade-offs involved in choosing between a longitudinal and a cross-sectional study design.

CONSIDERATIONS OF PRECISION Longitudinal studies give more precise estimates of temporal changes or treatment effects than cross-sectional studies of the same size. They achieve this gain in precision by eliminating

interindividual variability from the comparisons of interest. In a simple example, suppose that the variance of repeated measurements on the same individual is σ_w^2, and the variance of the average response among individuals is σ_a^2. Then the variance of the difference between measurements on two individuals is $2(\sigma_a^2 + \sigma_w^2)$, whereas the difference between two measurements of the same individual is $2\sigma_w^2$. If two treatments have different mean responses but the same variances, the sample sizes per treatment group, N_L and N_C, of longitudinal and cross-sectional studies of equal precision satisfy.

$$N_C = N_L\,[1 + (\sigma_a^2/\sigma_w^2)].$$

Longitudinal designs also vary in their efficiency. Machin (33) compared the efficiency of pure and mixed longitudinal designs for estimating a linear function of time. In the pure longitudinal study he assumed that the same n individuals were measured at p points in time. In his alternative design, m individuals were measured at k points, not necessarily the same for each individual, with $m < n$. This is a type of mixed longitudinal design in which, for example, not all individuals are measured at the same ages. Machin assumed that the correlation between measurements at any two successive points was ρ. He found that the relative efficiency of pure and mixed designs depends mainly on the value of ρ. If ρ is positive, as is usually the case with repeated measurements, the efficiency is greater for the mixed longitudinal design. Rao & Rao (42) also compared the efficiency of mixed longitudinal and cross-sectional designs. They found that the mixed design is preferable for moderate values of ρ if the objective is estimation of the difference in means for two occasions.

Besides increasing precision by eliminating between-individual variation, repeated measurements on the same individuals reduce bias due to differential selection or confounding. If we take differences between measurements on the same individual, the magnitude of the change cannot be attributed to differences in race, sex, or other individual characteristics. In this sense, repeated measures designs are analogous to matched studies in their approach to controlling external variables.

AGE-TIME-COHORT EFFECTS In the psychological and sociological literature, there has been much discussion of age, time, and cohort effects, and the roles of longitudinal and cross-sectional studies in estimating these effects (16, 23, 46, 49). The *age effect* represents changes in the average response due to the natural aging process. The *time of measurement effect* represents the impact of events in chronologic time that take place at the points of measurement. This includes the various treatments, external exposures, or changes in observers. The *cohort effect* represents past history specific to a particular cohort and contributes to all measurements on the

cohort. Examples may be events at birth, exposure to past epidemics or war, past environmental exposures, etc. It includes all time effects that took place before the start of the study. It also includes secular trends. The time of measurement effect is distinguished from the cohort effect in that it is a current temporary condition that affects all ages and cohorts. The three variables often actually serve as surrogates for other unobservable variables.

We can include all three effects in a simple additive model. Let Y_{ijk} be a measurement on a person from cohort i, of age j, and at time k. Then let

$$Y_{ijk} = c_i + a_j + t_k + e_{ijk}$$

where c_i, a_j, and t_k represent the cohort, age and time effects, respectively, and e_{ijk} is the error term. The three effects cannot be estimated simultaneously, since age and time determine the cohort. Because of this relationship, some authors, for example Goldstein (21), say that there are only two underlying dimensions and that no attempt should be made to distinguish three. Schaie (46) and van't Hof et al (49) point out that the three effects represent different biological processes, and that an effort to separate them should be made.

The study types differ in which of the three effects can be estimated and which are confounded. In the cross-sectional study, all measurements are made at a single time, so that no time effect can be estimated. In our model, t_k would be the same for all measurements. Also, we cannot separate the age and cohort effects in cross-sectional studies, since the age of the individual determines his or her cohort. Thus, cross-sectional studies yield biased estimates of the changes due to development or aging when cohort effects are substantial. In the pure longitudinal study we observe only one cohort, and comparisons within this group confound the age and time effects because the age uniquely determines the time of measurement.

Schaie (46) and van't Hof et al (49) advocate the mixed longitudinal design. They show that age, time, and cohort effects can be estimated provided that these effects are given special mathematical forms. For example, van't Hof et al model age effects as a quadratic function of time, and break up the time scale into three broad categories. They show how to use this design to estimate growth rates and norms.

Rao & Rao (42) observe that the results from the three types of studies will be similar only if the factors that influence growth are stable over a long period of time. In this situation, the time and cohort effects would be negligible. Guire & Kowalski (23) argue that although stability over time is usually not true for sociological studies, it is often true for studies of physical growth. In particular, the short-term time effect is usually assumed

to be zero, enabling us to use the pure longitudinal study to estimate age effects in a single cohort.

The problem of separating age, time, and cohort effects arises only when cohort is defined by year of birth. When cohort is defined by exposure or treatment group, as in many comparative studies, this cohort effect is not confounded with age and time effects and can be estimated separately. In comparing exposure groups, age, time, and year-of-birth effects can usually be controlled even when they are not separately estimable.

OTHER CONSIDERATIONS In some investigational settings, it is easy to choose between a longitudinal and a cross-sectional design. More often, the choice involves a series of trade-offs. Longitudinal designs give greater precision per observation, but observations may be more expensive or difficult to collect. Problems with missing or suspect data may be harder to solve in longitudinal studies. Implementation issues also influence design, since it is not always possible to sustain the commitment of investigators and participants, or the quality of study procedures. For these and other reasons, the relative advantages of cross-sectional and longitudinal designs have been debated for years (20, 35, 46, 49). Kowalski & Guire (31) provide an extensive bibliography on this topic. Here, we assume that the debate has, in a specific investigational setting, been resolved in favor of longitudinal studies, and we turn to a discussion of their design, conduct, and analysis.

Designing a Longitudinal Study

Once we have decided to conduct a longitudinal investigation, we must determine exactly how, when, where, and on whom the measurements will be taken. Longitudinal designs fundamentally have subjects crossed with occasions. This means that we must choose designs for selecting subjects and also the occasions at which they will be measured. Table 1 shows the data configuration for a balanced study in which each of N subjects is observed on the same p occasions. Because the issues involved in the design for subjects are not unique to longitudinal studies, we do not discuss them here. Cochran (9, 10) and Goldstein (21) examine some of the problems and methods for choosing these designs. In the sections that follow, we first look at designs for occasions, and then examine the further reduction of error variability.

DESIGNS FOR OCCASIONS In choosing a design for occasions, we must decide at which points to take the observations. In studies designed to compare the means for several occasions, the occasions must give the contrasts of interest. In observational studies, this depends primarily on the qualitative questions being examined. The corresponding issue in experi-

Table 1 Data configuration for a balanced longitudinal study

	Occasions			
Subjects	1	2	...	p
1	y_{11}	y_{12}	...	y_{1p}
2	y_{21}	y_{22}	...	y_{2p}
.....	.....	.....	.....	.....
N	y_{N1}	y_{N2}	...	y_{Np}

mental studies is the allocation of treatments over time; this has been studied extensively (7, 11).

The design for occasions is also important in growth studies in which development is modeled as a function of time or some other variable. How we define the measurement occasions can affect the efficiency of our study. Morrison (36) examined the optimal spacing of observations over intervals such as time. He considered constant, linear, and quadratic response functions, and covariances reflecting Weiner or stationary Markov processes. He found that equal spacing was either optimal or close to optimal for all situations considered.

Schlesselman (48) and Berry (3) examined design issues specifically for studies in which change was a function of time, giving emphasis to length of follow-up and frequency of measurement. Both authors considered a linear model for observations spaced equally in time, and analyzed the variance estimates of the estimated slope. Schlesselman computed an index of the precision of the slope estimate for various study durations and numbers of measurement points, and provided extensive tables showing the precision under the various designs. If $\hat{b}$ is the estimated slope of the line, then the standard error of this slope may be written $SE(\hat{b}) = w\sigma$, where σ is the standard deviation of the individual's points about the line. The w term is a constant, depending on the frequency and duration of the study. Schlesselman tabulated values of w. A small subset of Schlesselman's tables is shown in Table 2. We can see that both increased duration and number

Table 2 Values of w tabulated by duration and number of measurements

Number of measurements	Duration in years				
	1	3	5	10	20
2	1.414	0.471	0.283	0.141	0.071
6	1.195	0.398	0.239	0.120	0.060
12	0.920	0.307	0.184	0.092	0.046
24	0.678	0.226	0.136	0.068	0.034

of measurements lead to smaller standard errors, but that increased duration is more important in reducing variance.

The tabulated precision is for the slope for one individual. If the mean rate of change for a group is desired, an estimate of the variation between individuals is needed. The standard error of the average slope $\bar{b}$ for a group is $SE(\bar{b}) = [(\sigma_b^2 + w^2\sigma^2)/N]^{1/2}$, where σ_b^2 is the variance between individuals. Note that only the contribution of the variance about the individuals' lines, σ^2, is affected by w. This implies that frequence and duration can have only a limited impact on this standard error once $w\sigma \ll \sigma_b$. Any further reduction must be through increased sample size, N.

Berry gives similar estimates of precision, but considers the special case of measuring forced expiratory volume (FEV) in adult men. He obtains estimates of the between-individual, within-individual, and measurement error variances from several other studies of changes in FEV. The between-individual variability is actually the variability of the rate of linear decline between individuals, and the within-individual variability is calculated after removing the linear decline with time. In Schlesselman's notation, Berry assumes that $\sigma = 0.12$ liters and $\sigma_b = 0.04$ liters/year. Using these variances, Berry can give the true standard errors for the estimated linear decline in an individual and for a group for various follow-up durations and frequencies of measurement. Some of his values for the standard error of the average slope are given in Table 3. These are actually $(SE(\bar{b})\sqrt{N}$ since the standard error of the mean slope depends on the sample size. The standard errors are tabulated by years of follow-up and interval between measurements rather than number of measurements as in Table 2. Taking more measurements for the same duration has a small effect on the variance, while increasing duration reduces the variance substantially.

Decisions about sample size may be based on the same computations. For descriptive studies of one population, we can specify a desired precision and estimate the required sample size. For example, to achieve a standard error

Table 3 Standard errors (times $N^{1/2}$) for the mean slope of a group of individuals tabulated by duration and frequency of measurement

Interval between measurements	Duration in years			
	1	3	5	10
full duration	0.174	0.069	0.052	0.043
1 year	0.174	0.067	0.049	0.042
6 months	0.174	0.060	0.046	0.041
3 months	0.157	0.054	0.044	0.040
1 month	0.114	0.046	0.041	0.040

of ϵ for the average linear slope across individuals, $\bar{b}$, the necessary sample size is

$$N = \frac{1}{\epsilon^2}(\sigma_b^2 + w^2\sigma^2).$$

We need to specify duration and frequency as well as the variances σ^2 and σ_b^2 to arrive at N.

In comparative studies we wish to test differences between individuals. Schlesselman (47) gives an equation to calculate the sample size needed to detect a particular mean group difference Δ with given normal deviates Z_α and Z_β corresponding to the α and β error rates. This is

$$N = \frac{2(Z_\alpha + Z_\beta)^2}{\Delta^2}(\sigma_b^2 + w^2\sigma^2).$$

This assumes that we have already chosen the frequency and duration and have calculated w.

There are other issues, of course, in choosing the frequency and duration of a study. Berry suggests that missing data should be anticipated; more frequent measurements may be desirable to minimize the information lost. Schlesselman (48) has considered the possibility of nonlinear functional relationships, or more complex linear functions. He argues that the presence of these more complex functions does not preclude one from assuming a linear model in the design stage. Simple straight line functions can be used for simplicity in planning, even if other functions will be used in the analysis. Schlesselman cautions against using extreme designs, however, such as measurements taken only at the beginning and end of the study period, or very many measurements taken over a short time. More moderate designs will help us to detect nonlinearities in the data during analysis.

REDUCING ERROR VARIABILITY Once we have a particular design plan, we may still try to reduce error variability further. In any study there are many different sources of variability. We can identify three generic sources of error for repeated measurements on individuals:

1. *Measurement error* or *intraoccasion variability,* which we denote σ_e^2. This is the variability that would be evidenced if several measurements were taken on a single individual at the same occasion. It can be large, especially for measurements that are subjective or effort-dependent.
2. *Interoccasion variability,* which we denote σ_d^2. This is the variation between occasions in the "true" value for an individual, i.e. in the value that would be observed if there were no measurement error (for instance, a person's weight may fluctuate from day to day because of changes in diet or exercise).

3. *Variation among individuals,* which we denote as σ_a^2. If we think of each individual's values as fluctuating around some "true" or long-term average value, then these true values will vary among individuals. This is often the largest source of variability in serial measurements.

The variance of a difference in means is a function of the interoccasion and measurement errors. For instance, let N be the number of individuals, p be the number of occasions, and r be the number of observations per occasion for each individual. Then the variance of the difference in sample means for any two occasions is:

$$\frac{2}{N}(\sigma_d^2 + \frac{1}{r}\sigma_e^2). \qquad (*)$$

As discussed above, we have eliminated the between-individual variation σ_a^2 by taking differences of repeated measurements on individuals.

We can further reduce the variance (*) in two ways.

1. Reduce the component variances σ_d^2 and σ_e^2. To reduce interoccasion variability we may try to make experimental conditions as similar as possible on each occasion. For example, we could try to take weight or blood pressure measurements at the same time each day. To reduce σ_e^2 we can improve the accuracy of our equipment and procedures. This is discussed in the section on conduct.

2. Increase the size of the experiment. The variance (*) is a function of both N, the number of individuals, and r, the number of replications per occasion. As these get larger, the variance becomes smaller. If the measurement error σ_e^2 is negligible, however, as it is with some measures, taking more than one observation on each occasion is unnecessary. Of course, any precision gained through greater size must be weighed against the added cost of such increases.

CONDUCT

A study design is a plan for a data set that will efficiently test study hypotheses and estimate important parameters. Since analyses depend on the data actually collected, however, the value of a study depends equally on its design and its execution. Although the particulars of a well-conducted study depend on the variables measured and the goals of the study, some requirements for a well-executed study cut across disciplinary lines.

In any study involving measurement, sources of bias and measurement error must be controlled. Ordinarily, this requires extensive training of observers and evaluation of the validity and reliability of measurement instruments. Thus, in the Hypertension Detection and Follow-up Program

(27), a study in which blood pressure level was a critical measure of treatment effect, random zero sphygmomanometers were used to avoid bias arising from digit preference or other subjective factors. Blood pressure measurement devices were rigorously pretested, and observers were required to participate in a one week training course and pass a certification examination. These procedures were intended to minimize interobserver and interinstrument variability.

In a longitudinal study, there is an additional need to maintain study procedures over time. Instrument performance must be constantly monitored and systematic bias reestimated. Similarly, investigators should plan for training of new observers and recertification of established observers to avoid deterioration of measurement procedures. Because this element of a study does not contribute directly to study results, it is a natural candidate for reduced effort during periods of tight budgets. To avoid this mistake, study designs and study budgets should include adequate support for this work.

Investigators should establish a regular schedule for checking instruments and observers, and the results of these checks should be retained as a part of the study records. When appropriate, instrument and observer number should be recorded with each observation, so that these variables can be considered in the analysis if necessary. In some studies, instrument and observer variability are comparable in magnitude to the effects under study and may bias estimates of effects of interest if not carefully controlled.

Bias can also arise in the recording, transmission, and entry of study data. Although many studies emphasize extensive quality control for data entry to computer files, variability in reading participant records such as spirograms or coronary angiograms and inconsistency in coding questionnaire data often represent more important sources of variability. These sources of uncertainty can be quantified and controlled only by additional quality assurance activity in the data collection system. Although observer variability is a widely recognized phenomenon, specialists are often surprised by the extent of observer variability in interpreting diagnostic tests such as electrocardiograms. Thus, duplicate readings should be introduced to quantify this variability, and blinded reading should be required when there is potential for bias in comparative work.

A well-designed longitudinal study can also be threatened by problems of missing data. Data can be missing either because of procedural error during a regular visit or because a participant does not appear for a regular visit. Both events are harmful to the study. Most procedures for treating missing data in the analysis assume that data are missing at random (44), i.e. the probability of missing an observation does not depend on the value of that observation. When that assumption is true, the main consequences

of missing data are (*a*) inconvenience, because unbalanced data sets are more difficult to analyze, (*b*) loss of precision, because missing outcomes reduce the effective size of the study, and (*c*) problems with adjustment for covariates when their values are missing.

In longitudinal studies, the assumption that observations are missing at random is frequently unjustified. Participants who are lost to follow-up are often atypical in terms of mobility, social class, and general health. This is a special threat to comparative studies in which different groups have different follow-up procedures. This problem is managed by introducing extra procedures to maintain follow-up of study participants and to encourage participation at regular visits.

In summary, excellence in longitudinal research requires care in both the design and conduct of studies. The general goal in conducting the study is to execute the study as designed. Three important objectives in conducting longitudinal studies are the following:

1. Minimize and quantify instrument and observer variability.
2. Ensure accurate recording, coding, and transcription of data.
3. Minimize nonparticipation and obtain complete data at regular visits.

ANALYSIS

To analyze longitudinal data, one must specify the probability distribution for each subject's set of responses. We shall at first assume that the outcome variable is a measurement that has a normal (Gaussian) distribution, then discuss non-parametric methods and the analysis of categorical outcome variables at the end of this section. For normally distributed outcome variables, the probability distribution is completely specified by (*a*) the expected outcome on each occasion for each sample of subjects and (*b*) the variances and covariances of the several measurements of a single subject.

The approach to modeling the expected outcome depends on the goals of the study. In repeated measures studies, differences between occasions in the expected outcome for a single subject are attributed to changes in treatment or exposure conditions. The analysis begins by testing the equality of mean values over occasions and continues with estimation of the differences in means between occasions. The first part of the section on analysis shows how to use ANOVA techniques to perform that analysis. When changes in the expected outcome over occasions are due to growth or aging, or when occasions correspond to different levels of exposure, the analyst will want to model the changes over occasions. The second part of this section describes how to develop a model using polynomial or nonlinear growth curve analysis. We also explain the concept known as tracking.

In longitudinal studies, the unit of observation is the set or *vector* of observations for a single subject. In the balanced design represented by Table 1, this vector is denoted by $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})$. Even when we analyze a single outcome variable, longitudinal analysis frequently requires multivariate statistical methods. [For a discussion of the analysis of several outcome variables, see (31).] A model that does not recognize the interdependence among observations for a single subject can give seriously misleading results. If we assume that different measurements of a single subject are independent when they are actually positively correlated, the standard errors attributed to differences between means on different occasions will be too large, resulting in conservative tests and confidence intervals. For comparisons between groups, however, tests and confidence intervals will be nonconservative (26).

We represent the general multivariate normal distribution for $\mathbf{y}_i$ by writing

$$\mathbf{y}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

indicating that $\mathbf{y}_i$ has a multivariate normal distribution with mean value given by the vector $\boldsymbol{\mu}$ and variances and covariances given by the $p \times p$ matrix $\boldsymbol{\Sigma}$. Fortunately, we can sometimes make simplifying assumptions about the form of $\boldsymbol{\Sigma}$ that justify the use of univariate methods, as we illustrate in the next section.

Mixed-Model Analysis of Variance

Suppose that each subject in a single sample of size N is measured on the same p occasions characterized by different treatment or exposure conditions. The data set has the form shown in Table 1, and the unit of observation is the vector of p observations for a single subject.

We sometimes assume that observations on different occasions are independent except for a shared *subject effect,* a deviation from the occasion mean that is constant over occasions. This leads to the linear model

$$y_{ij} = \mu + \alpha_j + b_i + e_{ij},$$

where μ is the unknown, overall mean; α_j is the difference between the mean on occasion j and the overall mean; b_i is a contribution that varies among subjects and is common to all observations for a subject, e_{ij} *is a* random deviation that is independently distributed for different subjects or occasions. This is the mixed model of univariate ANOVA. When it is appropriate, the analyst can use univariate methods to investigate the occasion means. Since the model implies constant variance over occasions and constant variance between occasions, the adequacy of this model can be

evaluated by testing whether Σ has this special form using the likelihood ratio method (40, 51).

In repeated measures designs, the hypothesis often of interest is that the mean response is the same on each occasion. When the mixed model is correct and no observations are missing, this hypothesis can be tested by a ratio of mean squares having an F distribution with p-1 and $(N\text{-}1)(p\text{-}1)$ degrees of freedom (6). When the assumed covariance structure does not hold, this statistic has an F distribution with reduced degrees of freedom (22). ANOVA methods can also be used to estimate differences and other linear combinations of occasion means (6). For instance, if occasions correspond to different levels of an exposure variable, a one degree-of-freedom test for trend can be constructed.

MISSING DATA Although most articles and textbooks on the analysis of repeated measurements assume a complete data set, some observations are missing in most clinical and epidemiologic studies. When very few observations are missing, the analysis is minimally affected by omitting subjects with missing values or substituting estimates for missing values. Both of these methods can perform poorly, however, when the number of missing observations is substantial (24). In that situation, one needs an efficient method of analysis. Although many statisticians would recommend maximum likelihood methods for estimating and testing occasion means, the maximum likelihood estimates are often not available in closed algebraic form.

One popular strategy for computing the maximum likelihood estimates is to adapt one of the many computer programs that maximizes functions. These programs are often based on the Newton-Raphson algorithm or modifications thereof. Alternatively, Orchard & Woodbury (38) proposed a method for computing maximum likelihood estimates based on iterative reestimation of the missing values, and Dempster et al (12) unified this method, calling it the EM algorithm. Laird & Ware (32) have shown how to use the EM algorithm to fit a wide class of models for longitudinal data. Although this work has simplified the use of maximum likelihood methods in longitudinal studies to some extent, special computer programs are still required. The analyst working with longitudinal data containing missing values has to choose either to develop or adapt the computing software needed for optimal analysis or to use ad hoc methods with known limitations.

EXTENSIONS OF THE MIXED MODEL The mixed-model analysis may also be appropriate for more complex designs. For instance, the design on occasions may involve factorial arrangements of treatments. Similarly, several groups of subjects may be defined by different treatment sequences;

cross-over designs are a widely used example. These more complex designs can also be analyzed by univariate ANOVA methods as long as the assumption about error structure is justified. For an introduction to the extensive literature on the analysis of repeated-measures experiments using mixed-model methods, see Winer (51) and Bock (6).

Multivariate Analysis of Variance

When the data consist of the $N \times p$ array shown in Table 1, but the variances and covariances of the repeated measurements do not satisfy the assumptions of the mixed-model ANOVA, an alternative model is required. Although univariate methods can be developed if the covariance matrix has other special forms (37), validity of the analysis is ensured if one uses a general multivariate model. We write this model in univariate notation as

$$y_{ij} = \mu_j + e_{ij}$$

or in multivariate notation, as

$$\mathbf{y}_i = \mu + \mathbf{e}_i$$

where $\mathbf{e}_i$ is a $p \times 1$ vector having a normal distribution with mean $\mathbf{0}$ and arbitrary covariance matrix Σ.

ONE-SAMPLE PROBLEMS As in the mixed-model analysis, the objective of the multivariate analysis of a single sample is to characterize the occasion mean vector μ. If the occasions correspond to treatments, an analysis in terms of differences between occasions is appropriate. When the occasions correspond to points on a continuum, tests for trend or a polynomial representation may be useful. The previous remarks about missing data also apply to the multivariate analysis. Beale & Little (2) have described the implementation of the EM algorithm for maximum likelihood estimation in this setting.

PROFILE ANALYSIS The multivariate approach to analyzing several samples of individuals observed on the same occasions is sometimes called profile analysis. If the average response for each group is plotted at each time of measurement, these points and the lines connecting them form the group profile. Three questions can be asked about the profiles of different groups:

1. Are the profiles of the same shape (are the line segments parallel)?
2. If the profiles are parallel, are they at the same level?
3. If the profiles are parallel, are they horizontal?

Although the methods for this analysis are conceptually straightforward, they again are based on multivariate ANOVA (6, 51).

Growth Curve Models

In describing mixed-model and multivariate ANOVA for serial measurements, we implicitly assumed that differences between occasions were due to "treatment" effects or random variation. We turn now to another important application of longitudinal designs, the study of growth, development, and aging.

The methodology for analyzing serial measurements from such studies is known as growth curve analysis. Its objectives are (*a*) modeling of the temporal process of changing response and (*b*) investigating the effects of individual characteristics or experiences on that process.

When the period of observation is short or the pattern of change sufficiently simple, an individual's series of expected responses, or *growth curve,* can be described by a polynomial in time or its surrogate. The effects of individual characteristics can be expressed as changes in the polynomial coefficients, for instance the level, rate of change, or rate of acceleration of the response. This formulation of the growth curve problem permits analysis using the multivariate linear model.

Suppose once again that the data are obtained in the configuration shown in Table 1. To fit a polynomial to the mean growth curve, we could use multivariate ANOVA methods. It is more fruitful, however, to formulate a random effects model similar to our mixed model.

Suppose for specificity that each individual's growth curve is quadratic in time. We assume that the coefficients of this growth curve vary among individuals, and that deviations from this curve due to interoccasion variability and measurement error are independent with constant variance. We further assume that the polynomial coefficients are normally distributed in the population. We can write this model as

$$y_{ij} = a_i + b_i t_j + c_i t_j^2 + e_{ij},$$

where the e_{ij} are independent errors given the coefficients a_i, b_i, c_i, and

$$\begin{pmatrix} a_i \\ b_i \\ c_i \end{pmatrix} \sim N(\tau, \Lambda).$$

where τ is the vector of mean values of the polynomial coefficients and Λ the covariance matrix. One can also assume that the expected value of the polynomial coefficients depends on individual characteristics. These two-

stage models have been extensively studied (15, 40, 41) and methods for testing and estimation are well established. These models are especially useful for exploratory analysis, since the coefficients can be estimated for each individual, then explored further through graphical and unweighted regression analysis. For instance, the mean and variance of the slope, b, can be compared for groups with different exposure histories or other differences in experience.

In growth curve analysis, an unbalanced design can result either from missing data or from variation in measurement times among subjects. The random effects model extends to this setting (25), and the family of random effects models described by Laird & Ware (32) includes polynomial growth models with arbitrary patterns of observation times. In many situations, however, this iterative analysis is closely approximated by a very simple analysis in which we fit a polynomial growth curve for each individual, then analyze the effect of individual characteristics and time-invariant exposure variables on these coefficients by ordinary linear regression, using the coefficients as summary statistics. Especially when the patterns of observations are similar among individuals, we recommend this two-step analysis both for efficiency and ease of interpretation.

TRACKING Early studies of growth established that children tend to remain at a fixed percentile of the age-specific height distribution as they mature. This phenomenon, known as tracking, has interested both biological scientists and statisticians. The idea is especially important for potential risk factors for disease, such as blood pressure or serum cholesterol level. If these characteristics track into adult life, persons at high risk can be identified during childhood. A recent series of articles in *Biometrics* (17, 34, 50) described alternative models for this phenomenon. McMahan (34) proposed the model

$$y_{ij} = \mu_j + k_i\sigma_j + e_{ij},$$

where (μ_j, σ_j) are the mean and standard deviation of the response at time (or age) j, k_i is the deviation specific to individual i, and e_{ij} are independent errors. McMahan proposed an index measuring the fraction of interindividual variation explained by tracking. Dockery et al (14) applied this index to repeated measurements of height and forced expiratory volume in one second (FEV_1). They found that FEV_1 exhibited tracking comparable in strength to height, though this tracking is less apparent because FEV_1 is subject to large interoccasion variability.

Although McMahan's model captures the essential feature of tracking, his model does not lead naturally to comparative analyses. If interest cen-

ters on the effect of individual characteristics on growth, the preferred method begins by transforming the observations to achieve constant variance over occasions, and then uses the family of polynomial models.

Time Series Methods

A longitudinal data set can be viewed as a sample of short time series, one from each subject. Thus, an alternative to the general multivariate model may be suggested by consideration of the large family of models developed in the time series literature (1, 8). Although much of this literature is oriented to the analysis of a single, lengthy time series, some recent work has considered the analysis of numerous short series (18, 39). Autoregressive models often are the intuitively appealing approach to modeling the covariance pattern in serial measurements because they allow positive correlation between successive observations. For instance, in the one-sample problem without covariates, the first order autoregressive model implies that the deviations

$$y_{i,j} - \lambda y_{i,j-1}$$

are independent for different values of j. Kowalski & Guire (31) have described applications of time series methods to longitudinal data and Joreskog (29) has developed a powerful unified theory for time series analysis of longitudinal data. Because these models are no more efficient than multivariate models when the dataset is balanced, do not extend easily to highly unbalanced settings, and are relatively difficult to use for comparative analysis, they have not been widely used in comparative research.

Nonparametric Methods

The normality assumption required for the methods thus far described is not always reasonable. Some investigators have investigated nonparametric methods for growth curve analysis (19, 54), but they quickly become intractable or awkward to use when the design is unbalanced or individual characteristics are considered. Thus, when normality is in doubt, it is preferable to use a normalizing transformation of the data whenever possible. When the outcome is categorical, special methods are required. We discuss binary outcomes below.

Nonlinear Models

All of the analyses described thus far utilize linear models, that is, models in which each individual's measurements arise from a multivariate normal distribution and have expected values that depend linearly on exposure or treatment variables and individual characteristics. However, linear models

do not always suffice. Two important examples are (*a*) models for growth in stature and (*b*) models for binary outcome variables. We discuss these two problems in this section and illustrate the general issues in analyses using nonlinear models.

ANALYZING GROWTH IN STATURE In the 1930s, a number of longitudinal studies of childhood growth and development were initiated in the United States and Europe. Numeric scientists involved in these studies began to refine mathematical models for the shape of individual growth curves. Popular models included the Jenss (28) curve,

$$y(t) = a + bt - \exp(c+dt)$$

for growth from zero to six years, and the Gompertz curve,

$$y(t) = k \exp[-\exp(a-bt)]$$

suggested by Wright (53) and developed by Winsor (52) and Richards (43). The logistic model,

$$y(t) = \frac{k}{1 + \exp(a+bt)}$$

was also favored because of its simplicity, but none of these models performed well for the entire period from infancy to adult life. In an effort to describe the entire growth process, Bock and colleagues (5) proposed the double logistic model, and later the triple logistic model (4)

$$y(t) = \sum_{i=1}^{3} \frac{k_i}{1 + \exp(a_i+b_it)}$$

When these models are used to analyze serial measurements, analysts usually assume that the coefficients of the model vary over individuals. This is conceptually equivalent to the assumption of polynomial growth models with random coefficients, but the theory has not been comparably developed. Unified analysis of serial measurements on several individuals, each described by a nonlinear growth curve and with coefficients randomly distributed in the population, is technically difficult and requires simplifying assumptions and iterative methods (4).

A practical approach to this problem follows the unweighted analysis in the polynomial setting. To investigate how individual characteristics are related to growth, we can fit the selected nonlinear growth curve for each

respondent and summarize that individual's growth curve by the estimated coefficients of the nonlinear model. We then treat these estimated coefficients as independent and identically, normally distributed observations, with unknown covariance matrix and expected values depending linearly on the individual characteristics of interest. This dependence can then be investigated by standard multivariate linear regression methods. Once again, this method provides an intuitively appealing approach to data summarization and analysis. The method assumes that the individual estimates are approximately normally distributed, and that interindividual variation dominates intraindividual variation. The reasonableness of these assumptions increases with the number of observations on each individual.

ANALYZING SERIAL DICHOTOMOUS RESPONSE Korn & Whittemore (30) applied this idea successfully to the analysis of serial observations of the presence or absence of asthma symptoms on successive days for asthmatics participating in panel studies in the Los Angeles area. If, for individual i,

$$y_{it} = \begin{cases} 1, \text{ if symptoms on day } t \\ 0, \text{ otherwise,} \end{cases}$$

and $p_{it} = P(y_{it} = 1)$, Korn & Whittemore proposed the logistic regression model

$$\log\left(\frac{p_{it}}{1 - p_{it}}\right) = a_i + b_i y_{i,t-1} + \sum_{l=1}^{k} c_{il} x_{tl}.$$

In their model, the probability of symptoms depends on yesterday's status, $y_{i,t-1}$ and characteristics of day t such as air quality and weather variables $(x_{t1}, \ldots, x_{tk})$. The coefficients a_i, b_i, $c_{i1}, \ldots, c_{ik}$, are assumed to be normally distributed in the population, with covariance matrix **D.** If the maximum likelihood estimates, a_i, b_i, $c_{i1}, \ldots, c_{ik}$, of the logistic regression coefficients are computed for each individual, the asymptotic theory of maximum likelihood estimation implies that they will be approximately normally distributed with variance $\mathbf{D} + \Sigma$, where Σ is the conditional variance of the maximum likelihood estimates and can be estimated directly from the data. Korn & Whittmore used this approach to fit linear regression models for logistic regression coefficients on individual characteristics. Their analysis required iterative weighted least squares, with reestimation of the covariance matrix, **D**, at each iteration.

These examples suggest a general approach to analyzing interindividual variation in longitudinal data when observation of each individual is sufficiently extensive. Each individual's data can be summarized in terms of estimates of the several important parameters. At the second step, these estimates are treated as normally distributed with mean values dependent on individual characteristics, and variances expressed as a combination of between and within individual variance. The dependence of these coefficients on individual characteristics is then investigated through iteratively reweighted least squares regression. This general approach applies for both linear and nonlinear models for the serial observations of a single individual.

Computing Software

The most useful computing packages for performing the analyses described in this section are the BMDP series (13) and the Statistical Analysis System (45). Both packages have mixed-model ANOVA, multivariate ANOVA, and nonlinear regression programs. Although BMDP is slightly more flexible for this purpose, neither series can handle mixed models with large numbers of individuals, and most programs require balanced data sets with no missing observations. The LISREL program (29) is also quite flexible, especially for modeling serial correlation and latent variables. For approximate analysis of more complicated data sets, one can eliminate observations or fill in missing values to achieve balance. For growth curve analysis, the unweighted methods described previously can be used. However, optimal analysis requires likelihood-based methods. To use these methods, the analyst must adapt a function-maximizing program of the type provided in BMDP and several other packages, request a copy of programs used by previous investigators, or develop new software.

SUMMARY

Longitudinal studies have a long history in medical and social science research. They offer a natural approach to the study of development and aging that allows the separation of age and cohort effects. They can also be used to produce precise estimates of treatment contrasts not subject to between-individual variability. Yet they are often more difficult and entail greater expense per observation than cross-sectional studies. Thus the choice between a longitudinal and a cross-sectional design in a specific setting may require a careful statement of study goals and a comparison of the validity, precision, and feasibility of the two strategies.

The design of a longitudinal study has two aspects: a design for selecting subjects and a design for occasions. The issues involved in choosing subjects

are common to most observational studies, but the design for occasions involves issues unique to longitudinal studies. In studies of change over time, precision is influenced by sample size, frequency of measurement, and duration of measurement. Thus, the relative cost-effectiveness of different designs will depend on relative costs in these three dimensions. To protect the validity of a longitudinal study, investigators should plan monitoring procedures to standardize study procedures over time, and special effort may be required to reduce loss to follow-up.

There are several approaches to analyzing serial measurements. If the object is to compare means on the several occasions, a form of Analysis of Variance can be used, either univariate or multivariate, depending on whether simple assumptions about error structure hold and whether the design is balanced. For modeling the outcome as a function of time or some other variable, several forms of linear growth models can be used. Finally, if linear models are inappropriate, nonlinear models are available that, although more complex, may describe the pattern of growth more informatively.

Because they permit direct observation of temporal change in individuals, longitudinal studies often lead to refinements in models for development and aging. In settings in which cohort effects are important, longitudinal methods are essential for studying age-related changes. Thus, longitudinal studies will continue to play an important role in medical and social-scientific research, and, in some instances, represent the definitive method for studying temporal changes and the factors on which they depend.

ACKNOWLEDGMENTS

Research was supported by grants HLO7427 and GM29745 from the National Institutes of Health.

Literature Cited

1. Anderson, T. W. 1971. *Statistical Analysis of Time Series.* New York: Wiley. 374 pp.
2. Beale, E. M., Little, R. J. A. 1975. Missing values in multivariate analysis. *J. R. Statist. Soc. B* 37:129–45
3. Berry, G. 1974. Longitudinal observations. Their usefulness and limitations with special reference to the forced expiratory volume. *Bull. Physiol. Pathol. Resp.* 10:643–55
4. Bock, R. D., Thissen, D. M. 1976. Fitting multi-component models for growth in stature. *Proc. 9th Intl. Biometric Conf. Raleigh, NC* 1:431–42
5. Bock, R. D., Wainer, H., Petersen, A., Thissen, D., Murray, J., Roche, A. 1973. A parametrization for individual human growth curves. *Human Biol.* 45:63–80
6. Bock, R. D. 1975. *Multivariate Statistical Methods in Behavioral Research.* New York: McGraw-Hill. 623 pp.
7. Box, G. E. P., Hunter, W. G., Hunter, J. S. 1978. *Statistics for Experimenters.* New York: Wiley. 653 pp.
8. Box, G. E. P., Jenkins, G. M. 1970. *Time-Series Analysis: Forecasting and Control.* San Francisco: Holden-Day. 575 pp.
9. Cochran, W. G. 1963. *Sampling Techniques.* New York: Wiley. 403 pp.
10. Cochran, W. G. 1965. The planning of observational studies of human popula-

tions. *J. R. Statist. Soc. A* 128:234–66
11. Cochran, W. G., Cox, G. M. 1957. *Experimental Design.* New York: Wiley. 2nd ed.
12. Dempster, A., Laird, N. M., Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* 39:1–38
13. Dixon, W. J., Brown, M. B., eds. 1979. *Biomedical Computer Programs P-Series* Berkeley: Univ. Calif. Press. 880 pp.
14. Dockery, D. W., Berkey, C. S., Ware, J. H., Speizer, F. E., Ferris, B. G. Jr. 1982. Distribution of FVC and FEV_1 in children 6 to 11 years old. *Am. Rev. Resp. Dis.* Submitted
15. Fearn, T. 1975. A Bayesian approach to growth curves. *Biometrika* 62:89–100
16. Fienberg, S. E., Mason, W. M. 1979. Identification and estimation of age-period-cohort models in the analysis of discrete archival data. In *Sociological Methodology.* New York: Jossey-Bass
17. Foulkes, M. A., Davis, C. E. 1981. An index of tracking for longitudinal data. *Biometrics* 37:439–46
18. Fuller, W. A., Battese, G. E. 1974. Estimation of linear models with crossed-error structure. *J. Econometr.* 2:67–78
19. Ghosh, M., Grizzle, J. E., Sen, P. K. 1973. Nonparametric methods in longitudinal studies. *J. Am. Statist. Assoc.* 68:29–36
20. Goldfarb, N. 1960. *An Introduction to Longitudinal Statistical Analysis.* Glencoe, IL: Free Press. 220 pp.
21. Goldstein, H. 1979. *The Design and Analysis of Longitudinal Studies.* New York: Academic. 199 pp.
22. Greenhouse, S. W., Geisser, S. 1959. On methods in the analysis of profile data. *Psychometrika* 24:95–112
23. Guire, K. E., Kowalski, C. J. 1979. Mathematical description and representation of developmental change functions on the intra- and interindividual levels. In *Longitudinal Research in the Study of Behavior and Development,* ed. J. R. Nesselroade, P. B. Baltes, pp. 89–110. New York: Academic
24. Haitovsky, J. 1968. Missing data in regression analysis. *J. R. Statist. Soc. B* 30:67–82
25. Harville, D. A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* 72:320–40
26. Hoel, P. G. 1964. Methods for comparing growth type curves. *Biometrics* 20:859–72
27. Hypertension Detection and Follow-up Program Cooperative Group. 1976. The hypertension detection and follow-up program. *Prev. Med.* 5:207–15
28. Jenss, R. M., Bayley, N. 1937. A mathematical method for studying growth in children. *Human Biol.* 9:556–63
29. Joreskog, K. G. 1970. Estimation and testing of simplex models. *Br. J. Math. Stat. Psych.* 23:121–45
30. Korn, E. L., Whittemore, A. S. 1979. Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics* 35:795–802
31. Kowalski, C. J., Guire, K. E. 1974. Longitudinal data analysis. *Growth* 38:131–69
32. Laird, N. M., Ware, J. H. 1982. Random effects models for longitudinal studies. *Biometrics.* In press
33. Machin, D. 1975. On a design problem in growth studies. *Biometrics* 31:749–53
34. McMahan, C. A. 1981. An index of tracking. *Biometrics* 37:447–55
35. Monti, K. L., Koch, G. G., Sawyer, J. 1980. Segmented linear regression models applied to the analysis of data from a cross-sectional growth experiment. *Biomet. J.* 22:23–29
36. Morrison, D. F. 1970. The optimal spacing of repeated measurements. *Biometrics* 26:281–90
37. Morrison, D. F. 1976. *Multivariate Statistical Methods.* New York: McGraw-Hill. 415 pp. 2nd ed.
38. Orchard, T., Woodbury, M. A. 1972. A missing information principle: Theory and applications. *Proc. 6th Berkeley Symp. Math. Stat. Probab.* 1:697–715
39. Parks, R. W. 1967. Efficient estimation of a system of regression equations when disturbances are both serially and contemporaneously correlated. *J. Am. Statist. Assoc.* 62:500–9
40. Rao, C. R. 1965. The theory of least squares when parameters are stochastic and its application to the analysis of growth curves. *Biometrika* 52:447–58
41. Rao, C. R. 1975. Simultaneous estimation of parameters in different linear models and applications to biometric problems. *Biometrics* 31:545–54
42. Rao, M. N., Rao, C. R. 1966. Linked cross-sectional study for determining norms and growth rates: A pilot survey of Indian school-going boys. *Sankhya Ser. B* 28:237–58
43. Richards, F. J. 1959. A flexible growth function for empirical use. *J. Exp. Bot.* 10:290–300
44. Rubin, D. B. 1976. Inference and miss-

ing data (with discussion). *Biometrika* 63:581–92
45. SAS Inst. 1979. *SAS User's Guide, 1979 Edition.* Cary, NC: SAS Inst. Inc. 494 pp.
46. Schaie, K. W. 1965. A general model for the study of developmental problems. *Psychol. Bull.* 64:92–107
47. Schlesselman, J. J. 1973. Planning a longitudinal study. I. Sample size determination. *J. Chronic Dis.* 26:553–60
48. Schlesselman, J. J. 1973. Planning a longitudinal study. II. Frequency of measurement and study duration. *J. Chronic Dis.* 26:561–70
49. Van't Hof, M. A., Roede, M. J., Kowalski, C. J. 1977. A mixed longitudinal data analysis model. *Human Biol.* 49:165–79
50. Ware, J. H., Wu, M. C. 1981. Tracking: Prediction of future values from serial measurements. *Biometrics* 37:427–38
51. Winer, B. J. 1971. *Statistical Principles in Experimental Design.* New York: McGraw-Hill. 907 pp.
52. Winsor, C. P. 1932. The Gompertz curve as a growth curve. *Proc. Nat. Acad. Sci. USA* 18:1–8
53. Wright, S. 1926. Book review. *J. Am. Statist. Assoc.* 21:494
54. Zerbe, G. O., Walker, J. H. 1977. A randomization test for comparison of groups of growth curves with different polynomial design matrices. *Biometrics* 33:653–57