CHAPTER 1

Advances in longitudinal data analysis:
An historical perspective

Garrett Fitzmaurice and Geert Molenberghs

Contents
1.1 Introduction ................................................... ............................ 3
1.2 Early origins of linear models for longitudinal data analysis ...................... 3
1.3 Linear mixed-effects model for longitudinal data ..................................... 7
1.4 Models for non-Gaussian longitudinal data ........................................... 8
  1.4.1 Marginal or population-averaged models ....................................... 9
  1.4.2 Generalized linear mixed models ................................................ 16
  1.4.3 Conditional and transition models .............................................. 20
1.5 Concluding remarks ................................................................................ 21
Acknowledgments ......................................................................................... 22
References ...................................................................................................... 22

1.1 Introduction

There have been remarkable developments in statistical methodology for longitudinal data analysis in the past 25 to 30 years. Statisticians and empirical researchers now have access to an increasingly sophisticated toolbox of methods. As might be expected, there has been a lag between the recent developments that have appeared in the statistical journals and their widespread application to substantive problems. At least part of the reason why these advances have been somewhat slow to move into the mainstream is their limited implementation in widely available standard computer software. Recently, however, the introduction of new programs for analyzing multivariate and longitudinal data has made many of these methods far more accessible to statisticians and empirical researchers alike. Also, because statistical software is constantly evolving, we can anticipate that many of the more recent advances will soon be implemented. Thus, the outlook is bright that modern methods for longitudinal analysis will be applied more widely and across a broader spectrum of disciplines.

In this chapter, we take an historical perspective and review many of the key advances that have been made, especially in the past 30 years. Our review will be somewhat selective, and omissions are inevitable; our main goal is to highlight important and enduring developments in methodology. No attempt is made to assign priority to these methods. Our review will set the stage for the remaining chapters of the book, where the focus is on the current state of the art of longitudinal data analysis.

1.2 Early origins of linear models for longitudinal data analysis

The analysis of change is a fundamental component of so many research endeavors in almost every discipline. Many of the earliest statistical methods for the analysis of change were based on the analysis of variance (ANOVA) paradigm, as originally developed by
R. A. Fisher. One of the earliest methods proposed for analyzing longitudinal data was a mixed-effects ANOVA, with a single random subject effect. The inclusion of a random subject effect induced positive correlation among the repeated measurements on the same subject. Note that throughout this chapter we use the terms subjects and individuals interchangeably to refer to the participants in a longitudinal study. Interestingly, it was the British astronomer George Biddel Airy who laid the foundations for the linear mixed-model formulation (Airy, 1861), before it was put on a more formal theoretical footing in the seminal work of R. A. Fisher (see, for example, Fisher, 1918, 1925). Airy’s work on a model for errors of observation in astronomy predated Fisher’s more systematic study of related issues within the ANOVA paradigm (e.g., Fisher’s [1921, 1925] writings on the intraclass correlation). Scheffé (1956) provides a fascinating discussion of the early contributions of 19th century astronomers to the development of the theory of random-effects models. As such, it can be argued that statistical methods for the analysis of longitudinal data, in common with classical linear regression and the method of least squares, have their earliest origins in the field of astronomy.

The mixed-effects ANOVA model has a long history of use for analyzing longitudinal data, where it is often referred to as the univariate repeated-measures ANOVA. Statisticians recognized that a longitudinal data structure, with $N$ individuals and $n$ repeated measurements, has striking similarities to data collected in a randomized block design, or the closely related split-plot design. So it seemed natural to apply ANOVA methods developed for these designs (e.g., Yates, 1935; Scheffé, 1959) to the repeated-measure data collected from longitudinal studies. In doing so, the individuals in the study are regarded as the blocks or main plots. The univariate repeated-measures ANOVA model can be written as

$$Y_{ij} = X_{ij}'\beta + b_i + e_{ij}, \quad i = 1, \ldots, N; j = 1, \ldots, n,$$

where $Y_{ij}$ is the outcome of interest, $X_{ij}$ is a design vector, $\beta$ is a vector of regression parameters, $b_i \sim N(0, \sigma_b^2)$, and $e_{ij} \sim N(0, \sigma_e^2)$. In this model, the blocks or plot effects are regarded as random rather than fixed effects. The random effect, $b_i$, represents an aggregation of all the unobserved or unmeasured factors that make individuals respond differently. The consequence of including a single, individual-specific random effect is that it induces positive correlation among the repeated measurements, albeit with the following highly restrictive “compound symmetry” structure for the covariance: constant variance $\text{Var}(Y_{ij}) = \sigma_b^2 + \sigma_e^2$ and constant covariance $\text{Cov}(Y_{ij}, Y_{ik}) = \sigma_b^2$.

On the one hand, the univariate repeated-measures ANOVA model provided a natural generalization of Student’s (1908) paired t-test to handle more than two repeated measurements, in addition to various between-subject factors. On the other hand, it can be argued that this model was a Procrustean bed for longitudinal data because the blocks or plots were random rather than fixed by design and there is no sense in which measurement occasions can ever be randomized. Importantly, it is only when the within-subject factor is randomly allocated to individuals that randomization arguments can be made to justify the “compound symmetry” structure for the covariance. There is no basis for this randomization argument in the case of longitudinal data, where the within-subject factor is the measurement occasions. Recognizing that the compound symmetry assumption is restrictive, and to accommodate more general covariance structures for the repeated measures, Greenhouse and Geisser (1959) suggested a correction to the numerator and denominator degrees of freedom of tests derived from the univariate repeated-measures ANOVA (see also Huynh and Feldt, 1976).

In spite of its restrictive assumptions, and many obvious shortcomings, the univariate repeated-measures ANOVA model can be considered a forerunner of more versatile regression models for longitudinal data. As we will discuss later, the notion of allowing effects to vary randomly from one individual to another is the basis of many modern regression models.
EARLY ORIGINS OF LINEAR MODELS

for longitudinal data analysis. Also, it must be remembered that the elegant computational formulae for balanced designs meant that the calculation of ANOVA tables was relatively straightforward, albeit somewhat laborious. For balanced data, estimates of variance components could be readily obtained in closed form by equating ANOVA mean squares to their expectations; sometime later, Henderson (1963) developed a related approach for unbalanced data. So, from an historical perspective, an undoubted appeal of the repeated-measures ANOVA was that it was one of the few models that could realistically be fit to longitudinal data at a time when computing was in its infancy. This explains why, in those days, the key issue perceived to arise with incomplete data was lack of balance.

A related approach for the analysis of longitudinal data with an equally long history, but requiring somewhat more advanced computations, is the repeated-measures multivariate analysis of variance (MANOVA). While the univariate repeated-measures ANOVA is conceptualized as a model for a single response variable, allowing for positive correlation among the repeated measures on the same individual via the inclusion of a random subject effect, MANOVA is a model for multivariable responses. As originally developed, MANOVA was intended for the simultaneous analysis of a single measure of a multivariate vector of substantively distinct response variables. In contrast, while longitudinal data are multivariate, the vector of responses are commensurate, being repeated measures of the same response variable over time. So, although MANOVA was developed for multiple, but distinct, response variables, statisticians recognized that such data share a common feature with longitudinal data, namely, that they are correlated. This led to the development of a very specific variant of MANOVA, known as repeated-measures analysis by MANOVA (or sometimes referred to as multivariate repeated-measures ANOVA).

A special case of the repeated-measures analysis by MANOVA is a general approach known as profile analysis (Box, 1950; see also Geisser and Greenhouse, 1958; Greenhouse and Geisser, 1959). It proceeds by constructing a set of derived variables, based on a linear combination of the original sequence of repeated measures, and using relevant subsets of these to address questions about longitudinal change and its relation to between-subject factors. These derived variables provide information about the mean level of the response, averaged over all measurement occasions, and also about change in the response over time. For the most part, the primary interest in a longitudinal analysis is in the analysis of the latter derived variables. The multiple derived variables representing the effects of measurement occasions are then analyzed by MANOVA.

Box (1950) provided one of the earliest descriptions of this approach, proposing the construction of derived variables that represent polynomial contrasts of the measurement occasions; closely related work can be found in Danford, Hughes, and McNee (1960), Geisser (1963), Potthoff and Roy (1964), Cole and Grizzle (1966), and Grizzle and Allen (1969). Alternative transformations can be used, as the MANOVA test statistics are invariant to how change over time is characterized in the transformation of the original repeated measures. Although the MANOVA approach is computationally more demanding than the univariate repeated-measures ANOVA, an appealing feature of the method is that it allows assumptions on the structure of the covariance among repeated measures to be relaxed. In standard applications of the method, no explicit structure is assumed for the covariance among repeated measures (other than homogeneity of covariance across different individuals).

There is a final related approach to longitudinal data analysis based on the ANOVA paradigm that has a long history and remains in widespread use. In this approach, the sequence of repeated measures for each individual is reduced to a single summary value (or, in certain cases, a set of summary values). The major motivation behind the use of this approach is that, if the sequence of repeated measures can be reduced to a single number summary, then ANOVA methods (or, alternatively, non-parametric methods) for the analysis of a univariate response can be applied. For example, the area under the curve (AUC)
is one common measure that is frequently used to summarize the sequence of repeated
measures on any individual. The AUC, usually approximated by the area of the trapezoids
joining adjacent repeated measurements, can then be related to covariates (e.g., treatment or
intervention groups) using ANOVA. Wishart (1938) provided one of the earliest descriptions
of this approach in a paper with the almost unforgettable title “Growth-rate determinations
in nutrition studies with the bacon pig, and their analysis”; closely related methods can be
found in Box (1950) and Rao (1958).

Within a limited context, the three ANOVA-based approaches discussed thus far provided
the basis for a longitudinal analysis. However, all of these methods had shortcomings that
limited their usefulness in applications. The univariate repeated-measures ANOVA made
very restrictive assumptions about the covariance structure for repeated measures on the
same individual. The assumed compound symmetry form for the covariance is not appro-
priate for longitudinal data for at least two reasons. First, the constraint on the correlation
among repeated measurements is somewhat unappealing for longitudinal data, where the
correlations are expected to decay with increasing separation in time. Second, the assump-
tion of constant variance across time is often unrealistic. In many longitudinal studies the
variability of the response at the beginning of the study is discernibly different from the vari-
ability toward the completion of the study; this is especially the case when the first repeated
measurement represents a “baseline” response. Finally, as originally conceived, the repeated-
measures ANOVA model was developed for the analysis of data from designed experiments,
where the repeated measures are obtained at a set of occasions common to all individu-
als, the covariates are discrete factors (e.g., representing treatment group and time), and
the data are complete. As a result, early implementations of the repeated-measures ANOVA
could not be readily applied to longitudinal data that were irregularly spaced or incomplete,
or when it was of interest to include quantitative covariates in the analysis.

In contrast, the repeated-measures analysis by MANOVA did not make rest-
ictive as-
sumptions on the covariance among the longitudinal responses on the same individual. As
a result, the correlations could assume any pattern and the variability could change over
time. However, MANOVA had a number of features that also limited its usefulness. In
particular, the MANOVA formulation forced the within-subject covariates to be the same
for all individuals. There are at least two practical consequences of this constraint. First,
repeated-measures MANOVA cannot be used when the design is unbalanced over time (i.e.,
when the vectors of repeated measures are of different lengths and/or obtained at different
sequences of time). Second, the repeated-measures MANOVA (at least as implemented in
existing statistical software packages) did not allow for general missing-data patterns to
arise. Thus, if any individual has even a single missing response at any occasion, the entire
data vector from that individual must be excluded from the analysis. This so-called “list-
wise” deletion of missing data from the analysis often results in dramatically reduced sample
size and very inefficient use of the available data. Listwise deletion of missing data can also
produce biased estimators of change in the mean response over time when the so-called
“completers” (i.e., those with no missing data) are not a random sample from the target
population. Furthermore, balance between treatment groups is destroyed, hence the early
attraction of so-called imputation methods.

Finally, although the analysis of summary measures had a certain appeal due to the
simplicity of the method, it too had a number of distinct drawbacks. By definition, it forces
the data analyst to focus on only a single aspect of the repeated measures over time: when
n repeated measures are replaced by a single-number summary, there must necessarily be
some loss of information. Also, individuals with discernibly different response profiles can
produce the same summary measure. A second potential drawback is that the covariates
must be time-invariant; the method cannot be applied when covariates are time-varying.
Furthermore, many of the simple summary measures are not so well defined when there are
missing data or irregularly spaced repeated measures. Even in cases where the summary measure can be defined, the resulting analysis is not fully efficient. In particular, when some individuals have missing data or different numbers of repeated measures, the derived summary measures no longer have the same variance, thereby violating the fundamental assumption of homogeneity of variance for standard ANOVA models.

In summary, the origins of the statistical analysis of change can be traced back to the ANOVA paradigm. ANOVA methods have a long and extensive history of use in the analysis of longitudinal data. While ANOVA methods can provide a reasonable basis for a longitudinal analysis in cases where the study design is very simple, they have many shortcomings that have limited their usefulness in applications. In many longitudinal studies there is considerable variation among individuals in both the number and timing of measurements. The resulting data are highly unbalanced and not readily amenable to ANOVA methods developed for balanced designs. It was these features of longitudinal data that provided the impetus for statisticians to develop far more versatile techniques that can handle the commonly encountered problems of data that are unbalanced and incomplete, mistimed measurements, time-varying and time-invariant covariates, and responses that are discrete rather than continuous.

1.3 Linear mixed-effects model for longitudinal data

The linear mixed-effects model is probably the most widely used method for analyzing longitudinal data. Although the early development of mixed-effects models for hierarchical or clustered data can be traced back to the ANOVA paradigm (see, for example, Scheffé, 1959) and to the seminal paper by Harville (1977), their usefulness for analyzing longitudinal data, especially in the life sciences, was highlighted in the 1980s in a widely cited paper by Laird and Ware (1982). Goldstein (1979) is often seen as the counterpart for the humanities. The idea of allowing certain regression coefficients to vary randomly across individuals was also a recurring theme in the early contributions to growth curve analysis by Wishart (1938), Box (1950), Rao (1958), Potthoff and Roy (1964), and Grizzle and Allen (1969); these early contributions to growth curve modeling laid the foundation for the linear mixed-effects model. The idea of randomly varying regression coefficients was also a common thread in the so-called two-stage approach to analyzing longitudinal data. In the two-stage formulation, the repeated measurements on each individual are assumed to follow a regression model with distinct regression parameters for each individual. The distribution of these individual-specific regression parameters, or “random effects,” is modeled in the second stage. A version of the two-stage formulation was popularized by biostatisticians working at the U.S. National Institutes of Health (NIH). They proposed a method for analyzing repeated-measures data where, in the first stage, subject-specific regression coefficients are estimated using ordinary least-squares regression. In the second stage, the estimated regression coefficients are then analyzed as summary measures using standard parametric (or non-parametric) methods. Interestingly, this method for analyzing repeated-measures data became known as the “NIH method.” Although it is difficult to attribute the popularization of the NIH method to any single biostatistician at NIH, Sam Greenhouse, Max Halperin, and Jerry Cornfield introduced many biostatisticians to this technique. In the agricultural sciences, a similar approach was popularized in a highly cited paper by Rowell and Walters (1976). Rao (1965) put this two-stage approach on a more formal footing by specifying a parametric growth curve model that assumed normally distributed random growth curve parameters.

Although remarkably simple and useful, the two-stage formulation of the linear mixed-effects model introduced some unnecessary restrictions. Specifically, in the first stage, the covariates were restricted to be time-varying (with the exception of the column of 1s for
AN HISTORICAL PERSPECTIVE

the intercept); between-subject (or time-invariant) covariates could only be introduced in
the second stage, where the individual-specific regression coefficients were modeled as a
linear function of these covariates. The two-stage formulation placed unnecessary, and often
inconvenient, constraints on the choice of the design matrix for the fixed effects. But, from
an historical perspective, it provided motivation for the main ideas and concepts under-
lying linear mixed-effects models. The method can be viewed as based on summaries and
consequently it shares the disadvantages with such methods.

In the early 1980s, Laird and Ware (1982), drawing upon a general class of mixed mod-
els introduced earlier by Harville (1977), proposed a flexible class of linear mixed-effects
models for longitudinal data. These models could handle the complications of mistimed
and incomplete measurements in a very natural way. The linear mixed-effects model is given by

\[ Y_{ij} = X'_{ij} \beta + Z'_{ij} b_i + e_i \]

where \( Z_{ij} \) is a design vector for the random effects, \( b_i \sim N(0, G) \), and \( e_i \sim N(0, R_i) \).
Commonly, it is assumed that \( V_i = \sigma^2 I \), although additional correlation among the errors
can be accommodated by allowing more general covariance structures for \( V_i \) (e.g., autore-
gressive). In addition, alternative distributions for the random effects can be entertained.
The linear mixed-effects model proposed by Laird and Ware (1982) included the univari-
ate repeated-measures ANOVA and growth curve models for longitudinal data as special
cases. In addition, the Laird and Ware (1982) formulation of the model had two desir-
able features: first, there were fewer restrictions on the design matrices for the fixed and
random effects; second, the model parameters could be estimated efficiently via likelihood-
based methods. Previously, difficulties with estimation of mixed-effects models had held
back their widespread application to longitudinal data. Laird and Ware (1982) showed how
the expectation–maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) could
be used to fit this general class of models for longitudinal data. Soon after, Jennrich and
Schluchter (1986) proposed a variety of alternative algorithms, including Fisher scoring and
Newton–Raphson. Currently, maximum likelihood and restricted maximum likelihood esti-
mation, the latter devised to diminish the small-sample bias of maximum likelihood, are the
most frequently employed routes for estimation and inference (Verbeke and Molenberghs
2000; Fitzmaurice, Laird, and Ware, 2004).

So, by the mid-1980s, a very general class of linear models for longitudinal data had been
proposed that could handle issues of unbalanced data, due to either mistimed measurement
or missing data, could handle both time-varying and time-invariant covariates, and provided
a flexible, yet parsimonious, model for the covariance. Moreover, these developments ap-
peared at a time when there were great advances in computing power. It was not too long be-
fore these methods were available at the desktop and were being applied to longitudinal data
in a wide variety of disciplines. Nevertheless, many of the simple and simplifying procedures
stuck, out of habit and/or because they have become part of standard operating procedures.

1.4 Models for non-Gaussian longitudinal data

The advances in methods for longitudinal data analysis discussed so far have been based
on linear models for continuous responses that may be approximately normally distributed.
Next, we consider some of the parallel developments when the response variable is dis-
crete. The developments in methods for analyzing a continuous longitudinal response span
more than a century, from the early work on simple random-effects models by the British
astronomer Airy (1861) through the landmark paper on linear mixed-effects models for lon-
gitudinal data by Laird and Ware (1982). In contrast, many of the advances in methods for
discrete longitudinal data have been concentrated in the last 25 to 30 years, harnessing the
high-speed computing resources available at the desktop.
When the longitudinal response is discrete, linear models are no longer appropriate for relating changes in the mean response to covariates. Instead, statisticians have developed extensions of generalized linear models (Nelder and Wedderburn, 1972) for longitudinal data. Generalized linear models provide a unified class of models for regression analysis of independent observations of a discrete or continuous response. A characteristic feature of generalized linear models is that a suitable non-linear transformation of the mean response is assumed to be a linear function of the covariates. As we will discuss, this non-linearity raises some additional issues concerning the interpretation of the regression coefficients in models for longitudinal data. Statisticians have extended generalized linear models to handle longitudinal observations in a number of different ways; here we consider three broad, but quite distinct, classes of regression models for longitudinal data: (i) marginal or population-averaged models, (ii) random-effects or subject-specific models, and (iii) transition or response conditional models. These models differ not only in how the correlation among the repeated measures is accounted for, but also have regression parameters with discernibly different interpretations. These differences in interpretation reflect the different targets of inference of these models. Here we sketch some of the early developments of these models from an historical perspective; later chapters of this book will discuss many of these models in much greater detail. Because binary data are so common, we focus much of our review on models for longitudinal binary data. Most of the developments apply to, say, categorical data and counts equally well.

1.4.1 Marginal or population-averaged models

As mentioned above, the extensions of generalized linear models from the univariate to the multivariate response setting have followed a number of different research threads. In this section we consider an approach for extending generalized linear models to longitudinal data that leads to a class of regression models known as marginal or population-averaged models (see Chapter 3 of this volume). It must be admitted from the outset that the former term is potentially confusing; nonetheless it has endured faute de mieux. The term marginal in this context is used to emphasize that the model for the mean response at each occasion depends only on the covariates of interest, and not on any random effects or previous responses. This is in contrast to mixed-effects models, where the mean response depends not only on covariates but also on a vector of random effects, and to transition or generally conditional models (e.g., Markov models), where the mean response depends also on previous responses.

Marginal models provide a straightforward way to extend generalized linear models to longitudinal data. They directly model the mean response at each occasion, \( E(Y_{ij} | X_{ij}) \), using an appropriate link function. Because the focus is on the marginal mean and its dependence on the covariates, marginal models do not necessarily require full distributional assumptions for the vector of repeated responses, only a regression model for the mean response. As we will discuss later, this can be advantageous, as there are few tractable likelihoods for marginal models for discrete longitudinal data.

Typically, a marginal model for longitudinal data has the following three-part specification:

1. The mean of each response, \( E(\mu_{ij}) = \mu_{ij} \), is assumed to depend on the covariates through a known link function
   \[
   h^{-1}(\mu_{ij}) = X_{ij}'\beta.
   \]

2. The variance of each \( Y_{ij} \), given the covariates, is assumed to depend on the mean according to
   \[
   \text{Var}(Y_{ij} | X_{ij}) = \phi v(\mu_{ij}),
   \]
where $v(\mu_{ij})$ is a known variance function and $\phi$ is a scale parameter that may be known or may need to be estimated.

3. The conditional within-subject association among the vector of repeated responses, given the covariates, is assumed to be a function of an additional set of association parameters, $\alpha$ (and may also depend upon the means, $\mu_{ij}$).

Of the three, the first is the key component of a marginal model and specifies the model for the mean response at each occasion, $E(Y_{ij}|X_{ij})$, and its dependence on the covariates. However, there is an implicit assumption in the first component that is often overlooked. Marginal models assume that the conditional mean of the $j$th response, given $X_{i1}, \ldots, X_{in}$, depends only on $X_{ij}$, that is,

$$E(Y_{ij}|X_i) = E(Y_{ij}|X_{i1}, \ldots, X_{in}) = E(Y_{ij}|X_{ij}),$$

where obviously $X_i = (X_{i1}, \ldots, X_{in})$; see Fitzmaurice, Laird, and Rotnitzky (1993) and Pepe and Anderson (1994) for a discussion of the implications of this assumption. With time-invariant covariates, this assumption necessarily holds. Also, with time-varying covariates that are fixed by design of the study (e.g., time since baseline, treatment group indicator in a crossover trial), the assumption also holds, as values of the covariates are determined a priori by study design and in a manner unrelated to the longitudinal response. However, when a time-varying covariate varies randomly over time, the assumption may no longer hold. As a result, somewhat greater care is required when fitting marginal models with time-varying covariates that are not fixed by design of the study. This problem has long been recognized by econometricians (see, for example, Engle, Hendry, and Richard, 1983), and there is now an extensive statistical literature on this topic (see, for example, Robins, Greenland, and Hu, 1999).

The second component specifies the marginal variance at each occasion, with the choice of variance function depending upon the type of response. For balanced longitudinal designs, a separate scale parameter, $\phi_j$, can be specified at each occasion; alternatively, the scale parameter could depend on the times of measurement, with $\phi(t_{ij})$ being some parametric function of $t_{ij}$. Restriction to a single unknown parameter $\phi$ is especially limiting in the analysis of continuous responses where the variance of the repeated measurements is often not constant over the duration of the study.

The first two components of a marginal model specify the mean and variance of $Y_{ij}$, closely following the standard formulation of a generalized linear model. The only minor difference is that marginal models typically specify a common link function relating the vector of mean responses to the covariates. It is the third component that recognizes the characteristic lack of independence among longitudinal data by modeling the within-subject association among the repeated responses from the same individual. In describing this third component, we have been careful to avoid the use of the term correlation for two reasons. First, with a continuous response variable, the correlation is a natural measure of the linear dependence among the repeated responses and is variation independent of the mean response. However, this is not the case with discrete responses. With discrete responses, the correlations are constrained by the mean responses, and vice versa. The most extreme example of this arises when the response variable is binary. For binary responses, the correlations are heavily restricted to ranges that are determined by the means (or probabilities of success) of the responses. As a result, the correlation is not the most natural measure of within-subject association with discrete responses. For example, for two associated binary outcomes with probabilities of success equal to 0.2 and 0.8, the correlation can be no larger than 0.25.

Instead, the odds ratio is a preferable metric for association among pairs of binary responses. There are no restrictions for two outcomes, while they are mild for longer sequences of repeated measures. Second, for a continuous response that has a multivariate normal
distribution, the correlations, along with the variances and the means, completely specify the joint distribution of the vector of longitudinal responses. This is not the case with discrete data. The vector of means and the covariance matrix do not, in general, completely specify the joint distribution of discrete longitudinal responses. Instead, the joint distribution requires specification of pairwise and higher-order associations among the responses.

This three-part specification of a marginal model makes transparent the extension of generalized linear models to longitudinal data. The first two parts of the marginal model correspond to the standard generalized linear model, albeit with no explicit distributional assumptions about the responses. It is the third component, the incorporation of a model for the within-subject association among the repeated responses from the same individual, that represents the main extension of generalized linear models to longitudinal data. A crucial aspect of marginal models is that the mean response and within-subject association are modeled separately. This separation of the modeling of the mean response and the association among responses has important implications for interpretation of the regression parameters $\beta$. In particular, the regression parameters have population-averaged interpretations. They describe how the mean response in the population changes over time and how these changes are related to covariates. Note that the interpretation of $\beta$ is not altered in any way by the assumptions made about the nature or magnitude of the within-subject association.

From an historical perspective, it is difficult to pinpoint the origins of marginal models. In the case of linear models, the earliest approaches based on the ANOVA paradigm fit squarely within the framework of marginal models. In a certain sense, the necessity to distinguish marginal models from other classes of models becomes critical only for discrete responses. The development of marginal models for discrete longitudinal data has its origins in likelihood-based approaches, where the three-part specification given above is extended by making full distributional assumptions about the $n \times 1$ vector of responses, $Y_i = (Y_{i1}, \ldots, Y_{in})'$. Next, we trace some of these early developments and highlight many of the issues that have complicated the application of marginal models to discrete data, leading to the widespread use of alternative, semi-parametric methods.

At least three main research threads can be distinguished in the development of likelihood-based marginal models for discrete longitudinal data. Because binary data are so common, we focus much of this review on models for longitudinal binary data. One of the earliest likelihood-based approaches was proposed by Gumbel (1961), who posited a latent-variable model for multivariate binary data. In this approach, there is a vector of unobserved latent variables, say $L_{i1}, \ldots, L_{in}$, and each of these is related to the observed binary responses via

$$Y_{ij} = \begin{cases} 1 & L_{ij} \leq X_{ij}' \beta, \\ 0 & L_{ij} > X_{ij}' \beta. \end{cases}$$

Assuming a multivariate joint distribution for $L_{i1}, \ldots, L_{in}$ identifies the joint distribution for $Y_{i1}, \ldots, Y_{in}$, with

$$\Pr(Y_{i1} = 1, Y_{i2} = 1, \ldots, Y_{in} = 1) = \Pr(L_{i1} \leq X_{i1}' \beta, L_{i2} \leq X_{i2}' \beta, \ldots, L_{in} \leq X_{in}' \beta) = F(X_{i1}' \beta, X_{i2}' \beta, \ldots, X_{in}' \beta),$$

where $F(\cdot)$ denotes the joint cumulative distribution function of the latent variables. Furthermore, any dependence among the $L_{ij}$ induces dependence among the $Y_{ij}$. For example, a bivariate logistic distribution for any $L_{ij}$ and $L_{ik}$ induces marginally a logistic regression model for $Y_{ij}$ and $Y_{ik}$,

$$E(Y_{ij} | X_{ij}) = \frac{\exp(X_{ij}' \beta)}{1 + \exp(X_{ij}' \beta)},$$

with positive correlation between $Y_{ij}$ and $Y_{ik}$. 

MODELS FOR NON-GAUSSIAN LONGITUDINAL DATA

11
Although Gumbel’s (1961) model can accommodate more than two responses, the marginal covariance among the $Y_{ij}$ becomes quite complicated. Other multivariate distributions for the latent variables, with arbitrary marginals (e.g., logistic or probit) for the $Y_{ij}$ can be derived, but in general the joint distribution for $Y_{i1}, \ldots, Y_{in}$ is relatively complicated, as is the marginal covariance structure. As a result, these models were not widely adopted for the analysis of discrete longitudinal data. Closely related work, assuming a multivariate normal distribution for the latent variables, appeared in Ashford and Sowden (1970), Cox (1972), and Ochi and Prentice (1984). In the latter model, the $Y_{ij}$ marginally follow a probit model,

$E(Y_{ij}|X_{ij}) = \Phi(X'_{ij}\beta)$,

where $\Phi(\cdot)$ denotes the normal cumulative distribution function, and the model allows both positive and negative correlation among the repeated binary responses, depending on the sign of the correlation among the underlying latent variables. This model is often referred to as the “multivariate probit model.” Interestingly, the multivariate probit model can also be motivated through the introduction of random effects (see the discussion of generalized linear mixed models in Section 1.4.2).

One of the main drawbacks of the latent-variable model formulations that limited their application to longitudinal data is that they require $n$-dimensional integration over the joint distribution of the latent variables. In general, it can be computationally intensive to calculate or even approximate these integrals. In addition, the simple correlation structure assumed for the latent variables may be satisfactory for many types of clustered data but is somewhat less appealing for longitudinal data. In principle, however, a more complex covariance structure for the latent variables could be assumed.

At around the same time as Gumbel (1961) proposed his latent-variable formulation, a second approach to likelihood-based inferences was proposed by Bahadur (1961). Bahadur (1961) proposed an elegant expansion for an arbitrary probability mass function for a vector of responses $Y_{i1}, \ldots, Y_{in}$. The expansion for repeated binary responses is of the form

$f(y_{i1}, \ldots, y_{in}) = \left\{ \prod_{j=1}^{n} (\pi_{ij})^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}} \right\} \times \left\{ 1 + \sum_{j<k} \rho_{ijk} z_{ij} z_{ik} + \sum_{j<k<l} \rho_{ijkl} z_{ij} z_{ik} z_{il} + \cdots + \rho_{i1 \ldots n} z_{i1} \ldots z_{in} \right\}$,

where

$Z_{ij} = \frac{Y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1 - \pi_{ij})}}$,

$\pi_{ij} = E(Y_{ij})$, and $\rho_{ijk} = E(Z_{ij} Z_{ik})$, $\ldots$, $\rho_{i1 \ldots n} = E(Z_{i1} \ldots Z_{in})$. Here, $\rho_{ijk}$ is the pairwise or second-order correlation and the additional parameters relate to third- and higher-order correlations among the responses.

The Bahadur expansion has a particularly appealing property, shared with the multivariate probit model and many other marginal models, of being “reproducible” or “upwardly compatible” in the sense that the same model holds for any subset of the vector of responses. In addition, the multinomial probabilities for the vector of binary responses are relatively straightforward to obtain given the model parameters. Kupper and Haseman (1978) and Altham (1978) discussed applications of this model, albeit with very simple pairwise correlation structure and assuming higher-order terms are zero. The chief drawback of the Bahadur expansion that has limited its application to longitudinal data is its parameterization of the higher-order associations in terms of correlation parameters. As noted earlier, for discrete
MODELS FOR NON-GAUSSIAN LONGITUDINAL DATA

June 6, 2008 11:13 C6587 C6587\$C001

data there are severe restrictions on the correlations and dependence of the correlations on the means. Thus, for discrete data, the Bahadur model requires a complicated set of inequality constraints on the model parameters that make maximization of the likelihood very difficult. Except in very simple settings with a small number of repeated measures, the Bahadur model has not been widely applied to longitudinal data.

Because of the restrictions on the correlations, alternative multinomial models for the joint distribution of the vector of discrete responses have recently been proposed where the within-subject association is parameterized in terms of other metrics of association. For example, Dale (1984), McCullagh and Nelder (1989), Lipsitz, Laird, and Harrington (1990), Liang, Zeger, and Qaqish (1992), Becker and Balagtas (1993), Molenberghs and Lesaffre (1994), Lang and Agresti (1994), Glonek and McCullagh (1995), and others have proposed full likelihood approaches where the higher-order moments are parameterized in terms of marginal odds ratios. In closely related work, Ekholm (1991) parameterizes the association directly in terms of the higher-order marginal probabilities (see also Ekholm, Smith, and McDonald, 1995). An alternative approach is to parameterize the within-subject association in terms of conditional associations, leading to so-called “mixed-parameter” models (Fitzmaurice and Laird, 1993; Glonek, 1996; Molenberghs and Ritter, 1996). However, except in certain special cases (e.g., Markov models), these conditional association parameters have somewhat less appealing interpretations in the longitudinal setting; moreover, their interpretation is straightforward only in balanced longitudinal designs.

In virtually all of these later advances, the application of the methodology has been hampered by at least three main factors. First, unlike in the Bahadur model, there are no simple expressions for the joint probabilities in terms of the model parameters. This makes maximization of the likelihood somewhat difficult. Second, even with the current advances in computing, these models are difficult to fit except when the number of repeated measures is relatively small. Finally, many of these models are not robust to misspecification of the higher-order moments. That is, many of the likelihood-based methods require that the entire joint distribution be correctly specified. Thus, if the marginal model for the mean responses has been correctly specified but the model for any of the higher-order moments has not, then the maximum likelihood estimators of the marginal mean parameters will fail to converge in probability to the true mean parameters. The “mixed-parameter” models are an exception to the rule; however, even these models lose this robustness property when there are missing data.

A third approach to likelihood-based marginal models is to specify the entire multinomial distribution of the vector of repeated categorical responses and estimate the multinomial probabilities non-parametrically. This was the approach first proposed in Grizzle, Starmer, and Koch (1969). Specifically, they proposed a weighted least-squares (WLS) method for fitting a general family of models for categorical data; in recognition of its developers, the method is often referred to as the “GSK method.” Koch and Reinfurt (1971) and Koch et al. (1977) later recognized how these models could be applied to discrete longitudinal data; Stanish, Gillings, and Koch (1978), Stanish and Koch (1984), and Woolson and Clarke (1984) further developed the methodology for longitudinal analysis.

The GSK method provides a very general family of models for repeated categorical data, allowing non-linear link functions to relate the marginal expectations to covariates. The GSK method stratifies individuals according to values of the covariates and fully specifies the multinomial distribution of the vector of repeated categorical responses within each stratum. This method, for example, allows the fitting of logistic regression models to repeated binary data, albeit with the restrictions that the longitudinal study design be balanced on time, all covariates must be categorical, and there are sufficient numbers of individuals within covariate strata to estimate the multinomial probabilities non-parametrically as the sample proportions. The method requires the estimation of the covariance among the repeated
responses, within strata defined by covariate values; the covariance follows directly from the properties of the multinomial distribution. Asymptotically, the GSK method is equivalent to maximum likelihood estimation; thus, this approach was appealing for analyzing discrete longitudinal data when all of the conditions required for its use were met.

Although the GSK method was a landmark technique for the analysis of repeated categorical data, it had many restrictions that limited its usefulness. Specifically, it required that all covariates be categorical and sample sizes be of sufficient size to allow for stratification and separate estimation of the multinomial covariance in each covariate stratum. However, as the number of categorical covariates in the model increases, sparse data problems quickly arise due to Bellman’s (1961) “curse of dimensionality.” Furthermore, missing data are not easily handled by the GSK method because they require additional stratification by patterns of missingness. Thus, the GSK method was restricted to balanced designs with categorical covariates and relatively large sample sizes. It suffered from many of the same limitations as were noted for the repeated measures by MANOVA in Section 1.2.

Generally speaking, there have been a number of impediments to the application of likelihood-based marginal models for the analysis of discrete longitudinal data. The latent-variable model formulations, first proposed by Gumbel (1961), require high-dimensional integration over the joint distribution of the latent variables that is computationally too difficult. In contrast, methods that fully specify the multinomial probabilities for the response vector, such as the GSK method, are relatively straightforward to implement. However, the conditions for the use of the GSK method are typically not satisfied in many longitudinal settings. Alternative likelihood-based approaches that place more restrictions on the multinomial probabilities have proven to be substantially more difficult to implement. While the latter approaches do not require stratification and can incorporate a mixture of discrete and continuous covariates, they can be applied only in relatively simple cases.

For the most part, all of the likelihood-based approaches that have been proposed have been hampered by various combinations of the following factors. The first is the lack of a convenient joint distribution for discrete multivariate responses, with similar properties to the multivariate normal. Paradoxically, the joint distributions for discrete longitudinal data require specification of many higher-order moments despite the fact that there is, in some sense, substantially less information in the discrete than in the continuous data case. Second, with the exception of the Bahadur expansion, for many multinomial models there is no closed-form expression for the joint probabilities in terms of the model parameters. As such, there is no analog of the multivariate normal distribution for repeated categorical data that has simple and convenient properties. This makes maximization of the likelihood difficult. Third, all likelihood-based approaches face difficulties with sparseness of data once the number of repeated measures exceeds 5 or 6. Recall that a vector of repeated measures on a categorical response with $C$ categories requires specification of $C^m - 1$ multinomial probabilities. For example, with a binary response measured at 10 occasions, there are $2^{10} - 1$ (or 1023) non-redundant multinomial probabilities. So, while data on 200 subjects may be more than adequate for estimation of the marginal probabilities at each occasion, they can be wholly inadequate for estimation of the joint probabilities of the vector of responses due to the curse of dimensionality. Finally, many of these difficulties are compounded by the fact that likelihood-based estimates of the interest parameters, $\beta$, are quite sensitive to misspecification of the higher-order moments. Many of the proposed methods require that the entire joint distribution be correctly specified. For example, if the model for the mean response has been correctly specified but the model for the higher-order moments is incorrect, then the maximum likelihood estimator will fail to converge in probability to the true value of $\beta$. Although some of the computational difficulties previously mentioned can be ameliorated by faster and more powerful computers, many of the other problems reflect
the curse of dimensionality and cannot be easily handled with the typical amount of data collected in many longitudinal studies.

In the mid-1980s, remarkable advances in methodology for analyzing discrete longitudinal data were made when Liang and Zeger (1986) proposed the generalized estimating equations (GEE) approach. Because marginal models separately parameterize the model for the mean responses from the model for the within-subject association, Liang and Zeger (1986) recognized that it is possible to estimate the regression parameters in the former without making full distributional assumptions. The avoidance of distributional assumptions is potentially advantageous because, as we have discussed, there is no convenient and generally accepted specification of the joint multivariate distribution of $Y_i$ for marginal models when the responses are discrete. The appeal of the GEE approach is that it only requires specification of that part of the probability mechanism that is of scientific interest, the marginal means. By avoiding full distributional assumptions for $Y_i$, the GEE approach provided a remarkably convenient alternative to maximum likelihood estimation of multinomial models for repeated categorical data, without many of the inherent complications of the latter. Chapter 3 provides a comprehensive account of the GEE methodology.

The GEE approach advocated in Liang and Zeger (1986) was a natural extension of the quasi-likelihood approach (Wedderburn, 1974) for generalized linear models to the multivariate response setting, where an additional set of nuisance parameters for the within-subject association must be incorporated. The foundation for the GEE approach relied on the theory of optimal estimating functions developed by Godambe (1960) and Durbin (1960). Liang and Zeger (1986) highlighted how the GEE provides a unified approach to the formulation and fitting of generalized linear models to longitudinal and clustered data. They demonstrated the versatility of the GEE method in handling unbalanced data, mixtures of discrete and continuous covariates, and arbitrary patterns of missingness. Until the publication of their landmark paper (Liang and Zeger, 1986), methods for the analysis of discrete longitudinal data had lagged behind corresponding methods for continuous responses. Soon after, marginal models were being widely applied to address substantive questions about longitudinal change across a broad spectrum of disciplines. Their work also generated much additional theoretical and applied research on the use of this methodology for analyzing longitudinal data. For example, to improve upon efficiency, Prentice (1988) proposed joint estimating equations for both the main regression parameters, $\beta$, and the nuisance association parameters, $\alpha$.

The essential idea behind the GEE approach is to extend quasi-likelihood methods, originally developed for a univariate response, by incorporating additional nuisance parameters for the covariance matrix of the vector of responses. For example, given a model for the pairwise correlations, the corresponding covariance matrix can be constructed as the product of the standard deviations and correlations:

$$V_i = A_i^{1/2} \text{Corr}(Y_i) A_i^{1/2},$$

where $A_i$ is a diagonal matrix with $\text{Var}(Y_{ij}) = \phi v(\mu_{ij})$ along the diagonal, and $\text{Corr}(Y_i)$ is a correlation matrix (here a function of $\alpha$). In the GEE approach, $V_i$ is referred to as a “working” covariance matrix to distinguish it from the true underlying covariance matrix of $Y_i$. The term “working” in this context acknowledges uncertainty about the assumed model for the variances and within-subject associations. Because the GEE depend on both $\beta$ and $\alpha$, an iterative two-stage estimation procedure is required; this has been implemented in many widely available software packages. As noted by Crowder (1995), ambiguity concerning the definition of the working covariance matrix can, in certain cases, result in a breakdown of this estimation procedure.

In summary, the GEE approach has a number of appealing properties for estimation of the regression parameters in marginal models. First, in many longitudinal designs the GEE...
estimator of $\beta$ is almost efficient when compared to the maximum likelihood estimator. For example, it can be shown that the GEE has a similar expression to the likelihood equations for $\beta$ in a linear model for continuous responses that are assumed to have a multivariate normal distribution. The GEE also has an expression similar to the likelihood equations for $\beta$ in certain models for discrete longitudinal data. As a result, for many longitudinal designs, there is relatively little loss of precision when the GEE approach is adopted as an alternative to maximum likelihood. Second, the GEE estimator has a very appealing robustness property, yielding a consistent estimator of $\beta$ even if the within-subject associations among the repeated measures have been misspecified. It only requires that the model for the mean response be correct. This robustness property of GEE is important because the usual focus of a longitudinal study is on changes in the mean response. Although the GEE approach yields a consistent estimator of $\beta$ under misspecification of the within-subject associations, the usual standard errors obtained under the misspecified model for the within-subject association are not valid. However, valid standard errors for the resulting estimator $\hat{\beta}$ can be obtained using the empirical or so-called sandwich estimator of $\text{Cov}(\hat{\beta})$.

The sandwich estimator is also robust in the sense that, with sufficiently large samples, it provides valid standard errors when the assumed model for the covariances among the repeated measures is not correct.

1.4.2 Generalized linear mixed models

In the previous section, we discussed how marginal models can be considered an extension of generalized linear models that directly incorporate the within-subject association among the repeated measurements. In a certain sense, marginal models account for the consequences of the correlation among the repeated measures but do not provide any explanation for its potential source. An alternative approach for accounting for the within-subject association, and one that provides a source for the within-subject association, is via the introduction of random effects in the model for the mean response. Following the same basic ideas as in linear mixed-effects models, generalized linear models can be extended to longitudinal data by allowing a subset of the regression coefficients to vary randomly from one individual to another. These models are known as generalized linear mixed (effects) models (GLMMs), and they extend in a natural way the conceptual approach represented by the linear mixed-effects models discussed in Section 1.3; see Chapter 4 for a detailed overview. In GLMMs the model for the mean response is conditional upon both measured covariates and unobserved random effects; it is the inclusion of the latter that induces correlation among the repeated responses marginally, when averaged over the distribution of the random effects. However, as we discuss later, with non-linear link functions, the introduction of random effects has important ramifications for the interpretation of the “fixed-effects” regression parameters.

The generalized linear mixed model can be formulated using the following two-part specification:

1. Given a $q \times 1$ vector of random effects $b_i$, the $Y_{ij}$ are assumed to be conditionally independent and to have exponential family distributions with conditional mean depending upon both fixed and random effects,

$$h^{-1}\{E(Y_{ij}|b_i)\} = X'_{ij}\beta + Z'_{ij}b_i,$$

for some known link function, $h^{-1}(\cdot)$. The conditional variance is assumed to depend on the conditional mean according to $\text{Var}(Y_{ij}|b_i) = \phi v\{E(Y_{ij}|b_i)\}$, where $v(\cdot)$ is a known variance function and $\phi$ is a scale parameter that may be known or may need to be estimated.
MODELS FOR NON-GAUSSIAN LONGITUDINAL DATA

2. The random effects, \( b_i \), are assumed to be independent of the covariates, \( X_{ij} \), and to have a multivariate normal distribution, with zero mean and \( q \times q \) covariance matrix \( G \).

These two components completely specify a broad class of generalized linear mixed models. In principle, the conditional independence assumption in the first component is not necessary, but is commonly made. Similarly, any multivariate distribution can be assumed for the \( b_i \); in practice, however, it is common to assume that the \( b_i \) have a multivariate normal distribution.

Generalized linear mixed models have their foundation in simple random-effects models for binary and count data. The early literature on random-effects models for discrete data can be traced back to the development of random compounding models that introduced random effects on the response scale. For example, Greenwood and Yule (1920) introduced the negative binomial distribution as a compound Poisson distribution for count data, while Skellam (1948) provided an early discussion of the beta-binomial distribution for binary data. The beta-binomial model can be conceptualized as a two-stage model, where in the first stage the binary responses, \( Y_{i1}, \ldots, Y_{in} \), are assumed to be conditionally independent with common success probability \( p_i \), where \( \Pr(Y_{ij} = 1 | p_i) = E(Y_{ij} | p_i) = p_i \). In the second stage, the success probabilities, \( p_1, \ldots, p_N \), are assumed to be independently distributed with a beta density. The mean of the success probabilities can be related to covariates via an appropriate link function, such as a logit or probit link function. Although the beta-binomial model accounts for overdispersion relative to the usual binomial variance, the model is somewhat more natural for clustered rather than longitudinal data. As a result, the model has been used in a wide variety of different clustered data applications (e.g., Chatfield and Goodhardt, 1970; Griffiths, 1973; Williams, 1975; Kupper and Haseman, 1978; Crowder, 1978, 1979; Otake and Prentice, 1984; Aerts et al., 2002).

The main feature of the beta-binomial model that has limited its usefulness for analyzing longitudinal data is that it produces the same marginal distribution at each measurement occasion. While this may not be so problematic in certain clustered data settings (e.g., in study designs where \( X_{i1} = X_{i2} = \cdots = X_{in} \)), in a longitudinal study where interest is primarily in changes in the marginal means over time, this restriction on the marginal distributions is very unappealing. Nonetheless, the beta-binomial and other random compounding models motivated the later development of more versatile random-effects models. Recall that in the beta-binomial model, it is assumed that success probabilities vary randomly about a mean and the latter can be related to covariates via an appropriate link function, such as a logit link function. In contrast to this formulation, Pierce and Sands (1975) proposed an alternative model where the logit of \( p_i \) is assumed to vary about an expectation given by \( X_{ij}' \beta \),

\[
\text{logit}\{E(Y_{ij} | b_i)\} = X_{ij}' \beta + b_i,
\]

where \( b_i \) has a normal distribution with zero mean and constant variance. The appealing feature of the model proposed by Pierce and Sands (1975) is that the fixed and random effects are combined together on the same logistic scale. This model is often referred to as the simple “logit-normal model” and is very similar in spirit to the random intercept model for continuous outcomes discussed in Section 1.2. Although this model was remarkably simple, it proved to be difficult to fit at the time because maximum likelihood estimation required maximization of the marginal likelihood, averaged over the distribution of the random effect. This required integration, and no analytic solutions were available. The fact that the integral cannot be evaluated in a closed form limited the application of this model.

In closely related work, Ashford and Sowden (1970) proposed a very similar model, except with probit rather than logit link function. Interestingly, Ashford and Sowden’s (1970)
model with random intercept and probit link function was equivalent to the equicorrelated latent-variable model discussed in Section 1.4.1, leading to identical inferences provided the correlation is positive. Despite the fact that maximum likelihood estimation for even the simple logit-normal model was computationally demanding with the computer resources available at the time, Korn and Whittemore (1979) proposed a far more ambitious version of the model, where

\[ \logit\{E(Y_{ij}|b_i)\} = X'_{ij} \beta + Z'_{ij} b_i, \]

with \( Z_{ij} = X_{ij} \). Although their model was very general and avoided some of the obvious drawbacks of the simple logit-normal model, it was difficult to fit and required a very long sequence of repeated measures on each subject.

From an historical perspective, the papers by Ashford and Sowden (1970), Pierce and Sands (1975), and Korn and Whittemore (1979) laid the conceptual foundations for generalized linear mixed models; much of the work that followed focused on the thorny problem of estimation. In GLMMs the marginal likelihood is used as the basis for inferences for the fixed-effects parameters, complemented with empirical Bayes estimation for the random effects. In general, evaluation and maximization of the marginal likelihood for GLMMs requires integration over the distribution of the random effects. While this is, strictly speaking, true for the linear mixed-effects model as well, there the integration can be done analytically, so effectively a closed form for the marginal likelihood function arises, in which case the application of maximum or restricted maximum likelihood is straightforward. In the absence of an analytical solution, and because high-dimensional numerical integration can be very trying, a variety of approaches has been suggested for tackling this problem.

Because no simple analytic solutions were available, Stiratelli, Laird, and Ware (1984) proposed an approximate method of estimation for the logit-normal model, based on empirical Bayes ideas, that circumvented the need for numerical integration. Specifically, they avoided the need for numerical integration by approximating the integrands with simple expansions whose integrals have closed forms. The paper by Stiratelli, Laird, and Ware (1984) led to the development of a general approach for fitting GLMMs, known as penalized quasi-likelihood (PQL). Various authors (e.g., Schall, 1991; Breslow and Clayton, 1993; Wolfinger, 1993) motivated PQL as a Laplace approximation to the marginal likelihood for GLMMs. Despite the generality of this method, and its implementation in a variety of commercially available software packages, the PQL method can often yield quite biased estimators of the variance components, which in turn leads to biased estimators of \( \beta \), especially for longitudinal binary data. This motivated research on bias corrections (e.g., Breslow and Lin, 1995) and on more accurate approximations based on higher-order Laplace approximations (e.g., Raudenbush, Yang, and Yosef, 2000). In general, the inclusion of higher-order terms for PQL has been shown to improve estimation. Breslow and Clayton (1993) also considered an alternative approach, related to PQL, known as marginal quasi-likelihood (MQL). MQL differs from PQL by being based on an expansion around the current estimates of the fixed effects and around \( b_i = 0 \). In general, MQL yields severely biased estimators of the variance components, providing a good approximation only when the variance of the random effects is relatively small.

There has also been much recent research on alternative methods, including approaches based on numerical integration (e.g., adaptive Gaussian quadrature) and Markov chain Monte Carlo algorithms. In particular, adaptive Gaussian quadrature, with the numerical integration centered around the empirical Bayes estimates of the random effects, permits maximization of the marginal likelihood with any desired degree of accuracy (e.g., Anderson and Aitkin, 1985; Hedeker and Gibbons, 1994, 1996). Adaptive Gaussian quadrature is especially appealing for longitudinal data where the dimension of the random effects is often relatively low. Monte Carlo approaches to integration, for example Monte Carlo EM
MODELS FOR NON-GAUSSIAN LONGITUDINAL DATA

(McCulloch, 1997; Booth and Hobert, 1999) and Monte Carlo Newton–Raphson algorithms
(Kuk and Cheng, 1997), have been proposed. The hierarchical formulation of GLMMs also
makes Bayesian approaches quite appealing. For example, Zeger and Karim (1991) have
proposed the use of Monte Carlo integration, via Gibbs sampling, to calculate the posterior
distribution.

The normality assumption for the random effects in GLMMs leads, in general, to in-
tractable likelihood functions, except in the case of the linear mixed model for continuous
data. This is because the normal random-effects distribution is conjugate to the normal dis-
tribution for the outcome, conditional on the random effects. Lee and Nelder (1996, 2001,
2003) have extended this idea and propose using conjugate random-effects distributions in
contexts other than the classical normal linear model.

Finally, it is worth emphasizing some differences between GLMMs and the marginal mod-
el models discussed in Section 1.4.1. Although the introduction of random effects can simply be
thought of as a means of accounting for and explaining the potential sources of the corre-
lation among longitudinal responses, it has important implications for the interpretation of
the regression coefficients in GLMMs. The fixed effects, $\beta$, have somewhat different inter-
pretations than the corresponding regression parameters in marginal models. In GLMMs
the regression parameters have “subject-specific” interpretations. They represent the effects
of covariates on changes in an individual’s possibly transformed mean response per unit
change in the covariate, while controlling for all other covariates and the random effects.
This interpretation for $\beta$ can be better appreciated by considering the following example of
a simple logit-normal model given by

$$
\log \left\{ \frac{\Pr(Y_{ij} = 1|X_{ij}, b_i)}{\Pr(Y_{ij} = 0|X_{ij}, b_i)} \right\} = X'_{ij} \beta + b_i,
$$

where $b_i$ is assumed to have a univariate normal distribution with zero mean and constant
variance. The interpretation of a component of $\beta$, say $\beta_k$, is in terms of changes in any
given individual’s log odds of response for a unit change in the corresponding covariate, say
$X_{ijk}$. Because $\beta_k$ has interpretation that depends upon holding $b_i$ fixed, it is referred to
as a subject-specific effect. Note that this subject-specific interpretation of $\beta_k$ is far more
natural for a covariate that varies within an individual (i.e., a time-varying covariate). With
a time-invariant covariate, problems of interpretation arise because a change in the value of
the covariate also requires a change in the index $i$ of $X_{ijk}$ to, say, $X_{i'jk}$ (for $i \neq i'$). However,
$\beta_k$ then becomes confounded with differences between $b_i$ and $b_{i'}$. One way around this is to
think of the population, defined by all subjects sharing the same value of the random effect
$b_i$. The effect of a covariate is then conditional on changing $X_{ijk}$ within the fine population
defined by $b_i$.

In summary, the distinction between the regression parameters in GLMMs and marginal
models is best understood in terms of the targets of inference; a fuller discussion is given
in Chapter 7. In GLMMs, the target of inference is the individual because the regression
coefficients have interpretation in terms of contrasts of the transformed conditional means,

$$
E(Y_{ij} | X_{ij}, b_i).
$$

In contrast, in marginal models the target of inference is the population because the re-
gression parameters have interpretation in terms of contrasts of the transformed population
means,

$$
E(Y_{ij} | X_{ij}).
$$

For the special case of linear models, where an identity link function has been adopted, the
fixed effects in the model for the conditional means,

$$
E(Y_{ij} | X_{ij}, b_i) = X'_{ij} \beta + Z'_{ij} b_i,
$$
also happen to have interpretation in terms of the population means because

$$E(Y_{ij}|X_{ij}) = X'_{ij}\beta$$

when averaged over the distribution of the random effects. However, in general, for the non-linear link functions usually adopted for discrete data, this relationship no longer holds, and if

$$h^{-1}\{E(Y_{ij}|X_{ij}, b_i)\} = X'_{ij}\beta + Z'_{ij}b_i,$$

then

$$h^{-1}\{E(Y_{ij}|X_{ij})\} \neq X'_{ij}\beta$$

for any $\beta$.

### 1.4.3 Conditional and transition models

There is a third way in which generalized linear models can be extended to handle longitudinal data. This is accomplished by modeling the mean and time dependence simultaneously via conditioning an outcome on other outcomes or on a subset of other outcomes (see, for example, Molenberghs and Verbeke, 2005, Part III). A particular case is given by so-called transition, or Markov, models. Transition models are appealing due to the sequential nature of longitudinal data. In transition models, the conditional distribution of each response is expressed as an explicit function of the past responses and the covariates. Transition models can be considered conditional models in the sense of modeling the conditional distribution of the response at any occasion given the previous responses and the covariates. The dependence among the repeated measures is thought of as arising due to past values of the response influencing the present observation. In transition models, it is assumed that

$$h^{-1}\{E(Y_{ij}|X_{ij}, H_{ij})\} = X'_{ij}\beta + \sum_{r=1}^{s} \alpha_rf_r(H_{ij}), \quad (1.1)$$

where $H_{ij} = (Y_{i1}, \ldots, Y_{ij-1})$ denotes the history of the past responses at the $j$th occasion, and $f_r(H_{ij})$ denote some known functions (often, but not necessarily, linear functions) of the history of the past responses. For example, a first-order autoregressive, AR(1), generalized linear model is obtained when

$$\sum_{r=1}^{s} \alpha_rf_r(H_{ij}) = \alpha_1f_1(H_{ij}) = \alpha_1Y_{ij-1}.$$

A more general autoregressive model of order $s$, say, AR($s$), is obtained by incorporating the $s$ previously generated values of the response. In general, models where the conditional distribution of the response at the $j$th occasion, given $H_{ij}$, depends only on the $s$ immediately prior responses are known as Markov models of order $s$. When the response variable is discrete, these models are referred to as Markov chain models. With discrete data and non-identity link functions, it may be necessary to transform the history of the past responses, $H_{ij}$, in a manner similar to the transformation of the conditional mean, for example, $f_r(H_{ij}) = h^{-1}(H_{ij})$. Also, upon closer examination, the model given by (1.1) appears to make a strong assumption that the effects of the covariates are the same regardless of the actual history of past responses. However, this assumption can be relaxed by including interactions between relevant covariates and $H_{ij}$.

There is an extensive history to the use of Markov chains to model equally spaced discrete longitudinal data with a finite number of states or categories (e.g., Anderson and Goodman, 1957; Cox, 1958; Billingsley, 1961). In the simplest of models for longitudinal data, a first-order Markov chain, the transition probabilities are assumed to be the same for
each time interval. The resulting Markov chain can then be described in terms of the initial state and the set of transition probabilities. The transition probabilities are the conditional probabilities of going into each state, given the immediately preceding state. In a first-order Markov chain, there is dependence on the immediately preceding state but not on earlier outcomes. In the more general model given by (1.1), higher-order sequential dependence can be incorporated, with dependence on more than the immediately preceding state, and the transition probabilities can be allowed to vary over time. Moreover, the time dependence need not necessarily be a linear function of the history. Among others, Cox (1972), Korn and Whittemore (1979), Zeger, Liang, and Self (1985), and Ware, Lipsitz, and Speizer (1988) discuss transition models applicable to longitudinal data.

One appealing aspect of transition models is that the joint distribution of the vector of responses can be expressed as the product of a sequence of conditional distributions, that is,

\[
f(y_{i1}, \ldots, y_{in}; \beta, \alpha) = \prod_{j=1}^{n} f(y_{ij}|y_{i1}, \ldots, y_{i,j-1}; \beta, \alpha).
\]

Strictly speaking, for an \(s\)th-order Markov chain model, this is a conditional likelihood, given a set of \(s\) initial values. In the specification of the transition model given by (1.1), initial values of the responses are assumed to be incorporated into the covariates. In general, the unconditional distribution of the initial responses cannot be determined from the conditional distributions specified by (1.1). There are two ways to handle this initial value problem. The first is to treat the initial responses as a set of given constants rather than random variables and base estimation on the conditional likelihood, ignoring the contribution of the unconditional distribution of the initial responses. Maximization of the resulting likelihood is relatively straightforward; indeed, standard software for univariate generalized linear models can be used when \(f_r(H_{ij})\) does not depend on \(\beta\). Alternatively, the initial responses can be assigned the equilibrium distribution of the sequence of longitudinal responses. In general, the latter will yield more efficient estimates of the regression coefficient, \(\beta\).

Although Markov and autoregressive models have a long and extensive history of use for the analyses of time series data, their application to longitudinal data has been somewhat more limited. There are a number of features of transition models that limit their usefulness for the analysis of longitudinal data. In general, transition models have been developed for repeated measures that are equally separated in time; these models are more difficult to apply when there are missing data, mistimed measurements, and non-equidistant intervals between measurement occasions. In addition, estimation of the regression parameters \(\beta\) is very sensitive to assumptions concerning the time dependence; moreover, the interpretation of \(\beta\) changes with the order of the serial dependence. Finally, in many longitudinal studies \(\beta\) is not the usual target of inference because conditioning on the history of past responses may lead to attenuation of the effects of covariates of interest. That is, when a covariate is expected to influence the mean response at all occasions, its effect may be somewhat diminished if there is conditioning on the past history of the responses.

### 1.5 Concluding remarks

In the preceding sections, we traced the development of a very general and versatile class of linear mixed-effects models for longitudinal data when the response is continuous. These models can handle issues of unbalanced data, due to either mistimed measurement or missing data, time-varying and time-invariant covariates, and modeling of the covariance, in a flexible way. Linear mixed-effects models rely on assumptions of multivariate normality, and likelihood-based inferences for both the fixed and random effects are relatively straightforward. In contrast, when the longitudinal response is discrete, we have seen that there is more
than one way to extend generalized linear models to the longitudinal setting. This has led to the development of “marginal” and “conditional” models for non-Gaussian longitudinal data; in the former, there is no conditioning on past responses or random effects, while in the latter there is conditioning on either the response history or a set of random effects. Although this classification is useful for pedagogical purposes, it should be recognized that this distinction between classes of models is somewhat artificial and is made, in part, to emphasize certain aspects of interpretation that arise when analyzing discrete longitudinal data. In contrast to the situation for linear models, conditioning on past responses or random effects has important implications for the regression parameters in models for discrete longitudinal data. While their interpretation obviously changes in all cases, also the parameter estimates are different because in fact the underlying estimands cannot be compared directly. However, it is possible to combine features of these models, thereby blurring the distinctions. For example, Conaway (1989, 1990) has suggested extending mixed-effects models to include lagged responses, while, more recently, Heagerty and Zeger (2000) have developed “conditionally specified” marginal model formulations.

In general, we have seen that likelihood-based approaches are somewhat more difficult to formulate in the non-Gaussian data setting than is the case with continuous responses. This has led to various avenues of research where more tractable approximations have been developed (e.g., PQL methods) and where likelihood-based approaches have been abandoned altogether in favor of semi-parametric methods (e.g., GEE approaches).

Our review of the developments of regression models for longitudinal data has focused exclusively on extensions of generalized linear models. Limitations of space have precluded a discussion of non-linear models (i.e., models where the relationship between the mean and covariates is non-linear in the regression parameters) for longitudinal data; see Chapter 5 of this volume and Davidian and Giltinan (1995) for a comprehensive and unified treatment of this topic. Perhaps not surprisingly, the development of non-linear regression models for longitudinal data has faced many of the challenges and issues that were discussed in Section 1.4.2.

We conclude this chapter by noting that in almost every discipline there is increased awareness of the importance of longitudinal studies for studying change over time and the factors that influence change. This has led to a steady growth in the availability of longitudinal data, often arising from relatively complex study designs. The analysis of longitudinal data continues to pose many interesting methodological challenges and is likely to do so for the foreseeable future. The goal of the remaining chapters in this book is to highlight the current state of the art of longitudinal data analysis and to provide a glimpse of future directions.

Acknowledgments

This work was supported by the following grants from the U.S. National Institutes of Health: R01GM029745 and R01MH054693.

References


REFERENCES


REFERENCES


“Student” (Gosset, W. S.) (1908). The probable error of the mean. *Biometrika* 6, 1–25.


REFERENCES


