Advances in Analysis of Longitudinal Data

Robert D. Gibbons[1,2], **Donald Hedeker**[1,2], and **Stephen DuToit**[1,3]

Robert D. Gibbons: rdgib@uic.edu

[1] Center for Health Statistics, University of Illinois at Chicago, Illinois 60612

[2] Division of Epidemiology & Biostatistics, University of Illinois at Chicago, Illinois 60612

[3] Scientific Software International, Lincolnwood, Illinois 60712

Abstract

In this review, we explore recent developments in the area of linear and nonlinear generalized mixed-effects regression models and various alternatives, including generalized estimating equations for analysis of longitudinal data. Methods are described for continuous and normally distributed as well as categorical (binary, ordinal, nominal) and count (Poisson) variables. Extensions of the model to three and four levels of clustering, multivariate outcomes, and incorporation of design weights are also described. Linear and nonlinear models are illustrated using an example involving a study of the relationship between mood and smoking.

Keywords

OVERVIEW

Since the pioneering work of Laird & Ware (1982), statistical methods for the analysis of longitudinal data have advanced dramatically. Prior to this time, the standard approach to analysis of longitudinal data principally involved using the longitudinal data to impute end-points (e.g., last observation carried forward; LOCF) and then to simply discard the valuable intermediate time-point data, favoring the simplicity of analyses of change scores from baseline to study completion (or the last available measurement treated as if it was what would have been obtained had it been the end of the study), in some cases adjusting for baseline severity as well. Laird & Ware (1982) showed that generalized mixed-effects regression models could be used to perform a more complete analysis of all of the available longitudinal data under much more general assumptions regarding the missing data (i.e., missing at random; MAR). The net result was a more powerful set of statistical tools for analysis of longitudinal data that led to more powerful statistical hypothesis tests, more precise estimates of rates of change (and differential rates of change between experimental and control groups), and more general assumptions regarding missing data, for example because of study dropout. This early work has led to considerable related advances in statistical methodology for analysis of longitudinal data (for excellent reviews of this growing literature, see Diggle et al. 2002, Fitzmaurice et al. 2004, Goldstein 1995, Hedeker & Gibbons 2006, Longford 1993, Raudenbush & Bryk 2002, Singer & Willett 2003, and

Verbeke & Molenberghs 2000). Notable among these advances and relevant to this review are generalizations of the original Laird-Ware type model to the nonlinear case (relevant for analysis of binary, ordinal, nominal, count, and time-to-event outcomes), even more general forms of missing data (not missing at random; NMAR), higher levels of nesting such as three-level models (e.g., repeated observations nested within subjects and subjects nested within hospitals or clinics), alternative distributional assumptions for residual errors of measurement, correlated residual errors of measurement, multivariate mixed-effects regression models, and advances in sample size determination in the context of longitudinal studies. Computational advances in parameter estimation have also been seen, particularly in the area of nonlinear mixed effects regression models, where numerical evaluation of the likelihood function is more complex and requires high dimensional numerical integration (e.g., adaptive quadrature or Monte Carlo–type integration for full Bayes estimation of model parameters).

In the following sections, we provide a general and nontechnical overview of these recent advances and then illustrate some of their uses by applying them in analysis of several example datasets.

PROBLEMS INHERENT IN LONGITUDINAL DATA

Although longitudinal studies provide far more information than their cross-sectional counterparts and are therefore now in widespread use, they are not without limitations. In the following sections we review some of the major challenges associated with longitudinal data analysis.

Heterogeneity

Particularly in the behavioral sciences, individual differences are the norm rather than the exception. The overall mean response in a sample drawn from a population tells us little regarding the experience of the individual. In contrast to cross-sectional studies in which it is reasonable to assume that there are random fluctuations at each measurement occasion, when the same subjects are repeatedly measured over time, their responses are correlated over time, and their estimated trend line or curve can be expected to deviate systematically from the overall mean trend line. For example, behavioral and/or biological subject-level characteristics can increase the likelihood of a favorable response to a particular experimental intervention (e.g., a new pharmacologic treatment for depression), leading subjects with those characteristics to have a trend with higher slope (i.e., rate of change) than the overall average rate of change for the sample as a whole. In many cases, these personal characteristics may be unobservable, leading to unexplained heterogeneity in the population. Modeling this unobserved heterogeneity in terms of variance components that describe subject-level effects (or alternatively cluster-level effects such as classrooms, schools, clinics, etc.) is one way to accommodate the correlation of the repeated responses over time and to better describe individual differences in the statistical characterization of the observed data. These variance components are often termed "random effects," leading to terms like random-effects or mixed-effects regression models.

Correlated Errors of Measurement

In addition to heterogeneity in the population that leads to subject-specific deviations from the overall temporal response pattern, there is also often short-term correlated errors of measurement that are produced by the psychological state that a subject is in during measurement occasions that are close in time. This type of short-term residual correlation tends to decrease exponentially with the temporal distance between the measurement occasions. The addition of autocorrelated residuals (Chi & Reinsel 1989, Hedeker 1989) to

mixed-effects regression models allows for a more parsimonious analysis of the more subtle features of the longitudinal response process and results in more accurate estimates of uncertainty in parameter estimates, improved tests of hypotheses, and more accurate interval estimates.

Missing Data

Perhaps the most dramatic difficulty is the presence of missing data. Stated quite simply, not all subjects remain in the study for the entire length of the study. Reasons for discontinuing the study may be differentially related to the treatment. For example, some subjects may develop side effects to an otherwise effective treatment and must discontinue the study. Alternatively, some subjects might achieve the full benefit of the study early on and discontinue the study because they feel that their continued participation will provide no added benefit. The treatment of missing data in longitudinal studies is itself a vast literature, with major contributions by Laird (1988), Little (1995), Little & Rubin (2002), and Rubin (1976), to name a few. The basic problem is that even in a randomized and well-controlled clinical trial, the subjects who were initially enrolled in the study and randomized to the various treatment conditions may be quite different from the subjects who are available for analysis at the end of the trial. If subjects drop out because they already have derived full benefit from an effective treatment, an analysis that only considers those subjects who completed the trial may fail to show that the treatment was beneficial relative to the control condition. This type of analysis is often termed a "completer" analysis. To avoid this type of obvious bias, investigators often resort to an "intent-to-treat" analysis, in which the last available measurement is carried forward to the end of the study as if the subject had actually completed the study. This type of analysis, often termed an "end-point" analysis, introduces its own set of problems in that (*a*) all subjects are treated equally regardless of the actual intensity of their treatment over the course of the study, and (*b*) the actual response that would have been observed at the end of the study, if the subject had remained in the study until its conclusion, may in fact be quite different from the response made at the time of discontinuation. Returning to our example of the study in which subjects discontinue when they feel that they have received full treatment benefit, an endpoint analysis might miss the fact that some of these subjects may have had a relapse had they remained on treatment. Many other objections have been raised about these two simple approaches of handling missing data in longitudinal data, which have led to many more modern and far better motivated approaches to analysis of longitudinal data with missing observations.

Irregularly Spaced Measurement Occasions

It is not at all uncommon in real longitudinal studies, either in the context of designed experiments or naturalistic cohorts, for individuals to vary in the number of repeated measurements they contribute and even in the time at which the measurements are obtained. This may be due to dropout or simply due to different subjects having different schedules of availability. Although this can be quite problematic for traditional analysis-of-variance-based approaches (leading to highly unbalanced designs that can produce biased parameter estimates and tests of hypotheses), more modern statistical approaches to the analysis of longitudinal data are all but immune to the "unbalancedness" that is produced by having different times of measurement for different subjects. Indeed, this is one of the most useful features of the regression approach to this problem, namely the ability to use all of the available data from each subject, regardless of when it was specifically obtained.

Subjects Clustered in Centers

In addition to correlation produced by repeated measurements with the same individual, the clustering of individuals within ecological units (schools, classrooms, clinics, hospitals, counties, etc.) produces an additional source of correlation that violates the independence

assumption of traditional fixed-effects models. Simultaneous analysis of both clustered and longitudinal data leads to three-level (and even higher level) versions of the traditional two-level (repeated measurements nested with subjects) hierarchical or multilevel models. Although the magnitude of the variance component produced by clustering is typically much smaller than that produced by sampling repeated measurements within the same individual, ignoring this important source of variability leads to underestimates of standards errors of model parameters and tests of hypotheses with elevated Type I error rates (i.e., false positive rates). Three-level models are capable of simultaneously decomposing the overall variance into components related to within-subject effects and within-cluster effects. For example, a simple longitudinal study of patients within clinics may be based on the assumption that the intercept and slope of the linear time trend vary systematically from subject to subject, and the means of those time trends vary systematically from clinic to clinic.

STATISTICAL MODELS FOR ANALYSIS OF LONGITUDINAL AND/OR CLUSTERED DATA

In an attempt to provide a more general treatment of longitudinal data, with more realistic assumptions regarding the longitudinal response process and associated missing data mechanisms, statistical researchers have developed a wide variety of more rigorous approaches to the analysis of longitudinal data. Among these, the most widely used include mixed-effects regression models (Laird & Ware 1982) and generalized estimating equation (GEE) models (Zeger & Liang 1986). Variations of these models have been developed for both discrete and continuous outcomes and for a variety of missing data mechanisms. The primary distinction between the two general approaches is that mixed-effects models are full-likelihood methods and GEE models are partial-likelihood methods. The advantage of statistical models based on partial likelihood is that (*a*) they are computationally easier than full-likelihood methods, and (*b*) they generalize quite easily to a wide variety of outcome measures with quite different distributional forms. The price of this flexibility, however, is that partial likelihood methods are more restrictive in their assumptions regarding missing data than are their full-likelihood counterparts. In addition, full-likelihood methods provide estimates of person-specific effects (e.g., person-specific trend lines) that are quite useful in understanding interindividual variability in the longitudinal response process and in predicting future responses for a given subject or set of subjects from a particular subgroup (e.g., a county, a hospital, or a community). The distinctions between the various alternative statistical models for analysis of longitudinal data are fully explored in the following sections.

Mixed-Effects Regression Models

Generalized mixed-effects regression models are now quite widely used for analysis of longitudinal data. These models can be applied to both normally distributed continuous outcomes as well as categorical outcomes and other nonnormally distributed outcomes such as counts that have a Poisson distribution. Mixed-effects regression models are quite robust to missing data and irregularly spaced measurement occasions, and can easily handle both time-invariant and time-varying covariates. As such, they are among the most general of the methods for analysis of longitudinal data. As previously noted, they are sometimes called full-likelihood methods because they make full use of all available data from each subject. The advantage is that missing data are ignorable if the missing responses can be explained either by covariates in the model or by the available responses from a given subject. The disadvantage is that full-likelihood methods are more computationally complex than are quasi-likelihood methods, such as GEE.

Linear models—Traditional analysis of variance (ANOVA) methods, both univariate (repeated measures ANOVA) and multivariate (growth curve models), are of limited use because of restrictive assumptions concerning missing data across time and the variance-covariance structure of the repeated measures (Hedeker & Gibbons 2006, chapters 2 and 3). The univariate mixed-model or so-called repeated measures ANOVA assumes that the variances and covariances of the dependent variable across time are equal (i.e., compound symmetry). Alternatively, the multivariate analysis of variance for repeated measures only includes subjects with complete data across time. Also, these procedures focus on estimation of group mean trends across time and provide little help in understanding how specific individuals change across time. For these and other reasons, mixed-effects regression models (MRMs) have increasingly become popular for modeling longitudinal data.

Variants of MRMs have been developed under a variety of names: random-effects models (Laird & Ware 1982), variance component models (Dempster et al. 1981), multilevel models (Goldstein 1995), two-stage models (Bock 1989), random coefficient models (de Leeuw & Kreft 1986), mixed models (Longford 1987, Wolfinger 1993), empirical Bayes models (Hui & Berger 1983, Strenio et al. 1983), hierarchical linear models (Bryk & Raudenbush 1992), and random regression models (Bock 1983a,b; Gibbons et al. 1988). A basic characteristic of these models is the inclusion of random subject effects into the regression model in order to account for the influence of subjects on their repeated observations. These random subject effects thus describe each person's trend across time and explain the correlational structure of the longitudinal data. Additionally, they indicate the degree of subject variation that exists in the population of subjects.

Several features make MRMs especially useful in longitudinal research. First, subjects are not assumed to be measured on the same number of time-points; thus, subjects with incomplete data across time are included in the analysis. The ability to include subjects with incomplete data across time is an important advantage relative to procedures that require complete data across time because (*a*) by including all data, the analysis has increased statistical power, and (*b*) complete-case analysis may suffer from biases to the extent that subjects with complete data are not representative of the larger population of subjects. Because time is treated as a continuous variable in MRMs, subjects do not have to be measured at the same time-points. This is useful for analysis of longitudinal studies where follow-up times are not uniform across all subjects. Both time-invariant and time-varying covariates can be included in the model. Thus, changes in the outcome variable may be due to both stable characteristics of the subject (e.g., their gender or race) and characteristics that change across time (e.g., life events). Finally, whereas traditional approaches estimate average change (across time) in a population, MRMs can also estimate change for each subject. These estimates of individual change across time can be particularly useful in longitudinal studies where a proportion of subjects exhibit change across time that deviates from the average trend.

Consider the following simple linear mixed-effects regression model for the measurement *y* of individual *i* (*i* = 1, 2, … *N* subjects) on occasion *j* (*j* = 1, 2, … *n_j* occasions):

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij}.$$

Ignoring subscripts, this model represents the regression of the outcome variable *y* on the independent variable time (denoted *t*). The subscripts keep track of the particulars of the data, namely whose observation it is (subscript *i*) and when this observation was made (the subscript *j*). The independent variable *t* gives a value to the level of time and may represent

time in weeks, months, etc. Since y and t carry both i and j subscripts, both the outcome variable and the time variable are allowed to vary by individuals and occasions.

In linear regression models, the errors ε_{ij} are assumed to be normally and independently distributed in the population with zero mean and common variance σ^2. This independence assumption makes the typical general linear regression model unreasonable for longitudinal data. This is because the outcomes y are observed repeatedly from the same individuals, and so it is much more likely to assume that errors within an individual are correlated to some degree. Furthermore, the above model posits that the change across time is the same for all individuals since the model parameters (β_0, the intercept or initial level, and β_1, the linear change across time) do not vary by individuals. For both of these reasons, it is useful to add individual-specific effects into the model that will account for the data dependency and describe differential time trends for different individuals. This is precisely what MRMs do. The essential point is that MRMs therefore can be viewed as augmented linear regression models.

Random Intercept Model

A simple extension of the simple linear regression model is the random intercept model, which allows each subject to deviate from the overall mean response by a person-specific constant that applies equally over time:

$$y_{ij}=\beta_0+\beta_1 t_{ij}+\upsilon_{0i}+\varepsilon_{ij},$$

where υ_{0i} represents the influence of individual i on his/her repeated observations. Notice that if individuals have no influence on their repeated outcomes, then all of the υ_{0i} terms would equal 0. However, it is more likely that subjects will have positive or negative influences on their longitudinal data, and so the υ_{0i} terms will deviate from 0. Since individuals in a sample are typically thought to be representative of a larger population of individuals, the individual-specific effects υ_{0i} are treated as random effects. That is, υ_{0i} are considered to be representative of a distribution of individual effects in the population. The most common form for this population distribution is the normal distribution with mean 0 and variance σ_{υ}^2. In addition, the model assumes that the errors of measurement are conditionally independent, which implies that the errors of measurement are independent conditional on the random individual-specific effects υ_{0i}. Since the errors now have an influence due to individuals removed from them, this conditional independence assumption is much more reasonable than the ordinary independence assumption associated with the general linear model. The random intercept model is depicted graphically in Figure 1.

As mentioned, individuals deviate from the regression of y on t in a parallel manner in this model (since there is only one subject effect υ_{0i}). Thus, it is sometimes referred to as a random-intercepts model, with each υ_{0i} indicating how individual i deviates from the population trend. In this figure, the solid line represents the population average trend, which is based on β_0 and β_1. Also depicted are two individual trends, one below and one above the population (average) trend. For a given sample, there are N such lines, one for each individual. The variance term σ_{υ}^2 represents the spread of these lines. If σ_{υ}^2 is near zero, then the individual lines would not deviate much from the population trend. In this case, individuals do not exhibit much heterogeneity in their change across time. Alternatively, as individuals differ from the population trend, the lines move away from the population trend line and σ_{υ}^2 increases. In this case, there is more individual heterogeneity in time trends.

Random Intercept and Trend Model

For longitudinal data, the random intercept model is often too simplistic for a number of reasons. First, it is unlikely that the rate of change across time is the same for all individuals. It is more likely that individuals differ in their time trends; not everyone changes at the same rate. Furthermore, the compound symmetry assumption of the random intercept model is usually untenable for most longitudinal data. In general, measurements at points close in time tend to be more highly correlated than measurements further separated in time. Also, in many studies, subjects are more similar at baseline and grow at different rates across time. Thus, it is natural to expect that variability will increase over time.

For these reasons, a more realistic MRM allows both the intercept and time-trend to vary by individuals:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \upsilon_{0i} + \upsilon_{1i} t_{ij} + \varepsilon_{ij}.$$

In this model,

- β_0 is the overall population intercept,

- β_1 is the overall population slope,

- υ_{0i} is the intercept deviation for subject i, and

- υ_{1i} is the slope deviation for subject i.

As before, ε_{ij} is an independent error term distributed normally with mean 0 and variance σ^2.

The assumption regarding the independence of the errors is one of conditional independence, that is, they are independent conditional on υ_{0i} and υ_{1i}. With two random individual-specific effects, the population distribution of intercept and slope deviations is assumed to be bivariate normal $N(\mathbf{0}, \mathbf{\Sigma}_\upsilon)$, with the random-effects variance-covariance matrix given by

$$\sum_\upsilon = \left[\begin{array}{cc} \sigma^2_{\upsilon_0} & \sigma_{\upsilon_0 \upsilon_1} \\ \sigma_{\upsilon_0 \upsilon_1} & \sigma^2_{\upsilon_1} \end{array} \right].$$

This model can be thought of as a personal trend or change model since it represents the measurements of y as a function of time, both at the individual υ_{0i} and υ_{1i} and population β_0 and β_1 levels. The intercept parameters indicate the starting point, and the slope parameters indicate the degree of change over time. The population intercept and slope parameters represent the overall (population) trend, while the individual parameters express how subjects deviate from the population trend. Figure 2 represents this model graphically.

The interested reader is referred to Hedeker & Gibbons (2006, chapters 4–7) for a more detailed coverage of various linear mixed-effects regression models.

Non-linear models—Reflecting the usefulness of mixed-effects modeling and the importance of categorical outcomes in many areas of research, generalization of mixed-effects models for categorical outcomes has been an active area of statistical research. For dichotomous response data, several approaches adopting either a logistic or probit regression model and various methods for incorporating and estimating the influence of the random effects have been developed (Conaway 1989, Gibbons 1981, Gibbons & Bock 1987,

Goldstein 1991, Stiratelli et al. 1984, Wong & Mason 1985). We now describe a mixed-effects logistic regression model for the analysis of binary data. Extensions of this model for analysis of ordinal, nominal, and count data are described in detail by Hedeker & Gibbons (2006).

To formulate the logistic model, let p_i represent the probability of a positive outcome (i.e., Y_i = 1) for the ith individual. The probability of a negative outcome (i.e., $Y_i = 0$) is then $1 − p_i$. Denote the set of covariates as $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{i\,p})$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ is a $(p + 1) \times 1$ vector of corresponding regression coefficients. Then the logistic regression model is written as:

$$p_i = \Pr(Y_i = 1) = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})}$$

or

$$p_i = \frac{1}{1 + \exp(-\mathbf{x}_i'\boldsymbol{\beta})} = \psi(\mathbf{x}_i'\boldsymbol{\beta})$$

where $\psi(.)$ is the logistic cumulative distribution function (cdf), namely 1

$$\psi(z) = \frac{1}{1 + \exp(-z)}.$$

This model can also be represented in terms of the log odds or logit of the probabilities; namely,

$$\log \left[\frac{p_i}{1 - p_i} \right] = \mathbf{x}_i'\boldsymbol{\beta}$$

The numerator in the logit is the probability of a 1 response, and the denominator equals the probability of a 0 response. The ratio of these probabilities is the odds of a 1 response, and the log of this ratio is the log odds, or logit, of a 1 response. Notice that the log odds is equal to 0 when the probability of a 1 response equals 0.5 (i.e., equal odds of a response in category 0 and 1), is negative when the probability is less than 0.5 (i.e., odds favoring a response in category 0), and is positive when the probability is greater than 0.5 (i.e., odds favoring a response in category 1).

In logistic regression, the logit is called the link function because it maps the (0,1) range of probabilities unto the $(−\infty, \infty)$ range of linear predictors. The logit is linear in its parameter vector and so has many of the desirable properties of a linear regression model, albeit in terms of the logits. Note, however, that the logistic regression model is linear in terms of the logits and not the probabilities. In terms of the probabilities, the logistic regression model posits an s-shaped logistic relationship between the values of x and the probabilities as illustrated in Figure 3.

In contrast, the relationship between x and the logit is linear as shown in Figure 4.

Since the model is linear in terms of the logits, interpretation of the parameters of the logistic regression model is in terms of the logits. Thus, the intercept β_0 is the log odds of a positive outcome for an individual with a set of covariates $\mathbf{x}_i = 0$, and β_p measures the change in the log odds for a unit change in x_p holding all other covariates constant. Often the regression coefficients in logistic regression models are expressed in exponential form, namely $\exp(\beta_p)$. This transformation yields an odds ratio interpretation for the regressors, namely the ratio of the odds of a positive response for a unit change in x.

When the binary responses are clustered, for example repeatedly measured within individuals, or clustered within clinics, schools, or other social or ecological strata, the fixed-effects logistic regression model fails in its assumptions to accurately characterize the dependency in the data. Basically, the fixed-effects model assumes that the observations are independent, which they clearly are not when they are clustered within individuals. As described previously for the linear mixed-effects regression model, one solution to this problem is to generalize the model to the case of a combination of fixed (e.g., treatment) and random (e.g., time-trend coefficients) effects. The random effects allow the correlation between the clustered (e.g., repeated) measurements to be incorporated into the estimates of parameters, standard errors, interval estimates, and tests of hypotheses. One can conceptualize the random effects as representing subject-specific differences in the propensity to respond over time. Those subjects with higher response propensity will exhibit an increased probability of a positive response, conditional on their values of the fixed-effects (e.g., treatment group) included in the model.

To set the notation, let i denote the level-2 units (individuals) and let j denote the level-1 units (nested observations). Assume that there are $i = 1, \ldots, N$ level-2 units and $j = 1, \ldots, n_i$ level-1 units nested within each level-2 unit. The total number of level-1 observations across level-2 units is given by $n = \sum_{i=1}^{n} n_i$. Let y_{ij} be the value of the dichotomous outcome variable, coded 0 or 1, associated with level-1 unit j nested within level-2 unit i. The logistic regression model is written in terms of the log odds (i.e., the logit) of the probability of a response, denoted p_{ij}. Considering first a random-intercept model, augmenting the logistic regression model with a single random effect yields:

$$\log\left[\frac{p_{ij}}{1 - p_{ij}}\right] = \mathbf{x}_i'\boldsymbol{\beta} + \upsilon_i,$$

where $\mathbf{x}_{ij}$ is the $(p + 1) \times 1$ covariate vector (includes a 1 for the intercept), $\boldsymbol{\beta}$ is the $(p + 1) \times 1$ vector of unknown regression parameters, and υ is the random subject effect (one for each level-2 subject). These are assumed to be distributed in the population as $N(0, \sigma_\upsilon^2)$. For convenience and computational simplicity, in models for categorical outcomes the random effects are typically expressed in standardized form. For this, $\upsilon_i = \sigma_\upsilon \theta_i$ and the model is given as:

$$\log\left[\frac{p_i}{1 - p_i}\right] = \mathbf{x}_{ij}'\boldsymbol{\beta} + \sigma_\upsilon \theta_i.$$

Notice that the random-effects variance term (i.e., the population standard deviation σ_υ is now explicitly included in the regression model). Thus, it and the regression coefficients are on the same scale, namely, in terms of the log-odds of a response. In terms of the underlying latent y, the model is written as:

$$\log\left[\frac{p_i}{1-p_i}\right] = \mathbf{x}'_{ij}\boldsymbol{\beta} + \sigma_\upsilon\theta_i + \varepsilon_{ij}.$$

This representation helps to explain why the regression coefficients from a mixed-effects logistic regression model do not typically agree with those obtained from a fixed-effects logistic regression model, or for that matter from a GEE logistic regression model which has regression coefficients that agree in scale with the fixed-effects model. In the mixed model, the conditional variance of the latent y given $\mathbf{x}$ equals $\sigma_\upsilon^2 + \sigma_\varepsilon^2$, whereas in the fixed-effects model this conditional variance equals only the latter term σ_ε^2 (which equals either $\pi^2/3$ or 1 depending on whether it is a logistic or probit regression model, respectively). As a result, equating the variances of the latent y under these two scenarios yields:

$$\beta_M = \sqrt{\frac{\sigma_\upsilon^2 + \sigma_\varepsilon^2}{\sigma_\varepsilon^2}}$$

where β_F and β_M represent the regression coefficients from the fixed-effects and (random-intercepts) mixed-effects models, respectively. In practice, Zeger et al. (1988) have found that $(15/16)^2\pi^2/3$ works better than $\pi^2/3$ for σ_ε^2 in equating results of logistic regression models.

Several authors have commented on the difference in scale and interpretation of the regression coefficients in mixed models and marginal models, like the fixed-effects and GEE models (Neuhaus 1991, Zeger et al. 1988). Regression estimates from the mixed model have been termed "subject specific" to reinforce the notion that they are conditional estimates, conditional on the random (subject) effect. Thus, they represent the effect of a regressor on the outcome controlling for or holding constant the value of the random subject effect. Alternatively, the estimates from the fixed-effects and GEE models are "marginal" or "population-averaged" estimates that indicate the effect of a regressor averaging over the population of subjects. This difference of scale and interpretation only occurs for nonlinear regression models like the logistic regression model. For the linear model, this difference does not exist.

The model can be easily extended to include multiple random effects. For this, denote $\mathbf{z}_{ij}$ as the $r \times 1$ vector of random-effect variables (a column of ones is usually included for the random intercept). The vector of random effects $\boldsymbol{\upsilon_i}$ is assumed to follow a multivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma_\upsilon}$. To standardize the multiple random effects $\boldsymbol{\upsilon_i} = \mathbf{T}\boldsymbol{\theta_i}$ where $\mathbf{TT}' = \boldsymbol{\Sigma_\upsilon}$ is the Cholesky decomposition of $\boldsymbol{\Sigma_\upsilon}$. The model is now written as:

$$\log\left[\frac{p_{ij}}{1-p_{ij}}\right] = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{T}\boldsymbol{\theta_i}.$$

As a result of the transformation, the Cholesky factor $\mathbf{T}$ is usually estimated instead of the variance-covariance matrix $\boldsymbol{\Sigma_\upsilon}$. As the Cholesky factor is essentially the matrix square-root of the variance-covariance matrix, this allows more stable estimation of near-zero variance terms.

Two-level versus three-level models—Traditionally, mixed-effects regression models, both linear and nonlinear, are two-level models; for example, subjects repeatedly measured over time. Gibbons & Hedeker (1997) generalized the two-level mixed-effects logistic regression model to the case of three-level data (e.g., subjects repeatedly measured over time and clustered within clinics), and Hedeker & Gibbons (2006) reviewed a wide variety of linear and nonlinear three-level mixed-effects regression models and their application.

The amount of dependency in the data that is observable due to the clustering of the data is measured by the intraclass correlation. When the intraclass correlation equals 0, there is no association among subjects from the same cluster and analysis which ignores the clustering of the data is valid. For certain variables, however, intraclass correlation levels have been observed between 5% and 12% for data from spouse pairs and between 0.05% and 0.85% for data clustered by counties. As the intraclass correlation increases, the amount of independent information from the data decreases, inflating the Type I error rate of an analysis that ignores this correlation. Thus, statistical analysis that treats all subjects as independent observations may yield tests of significance that are too liberal.

The mixed model can be augmented to handle the clustering of data within centers. For this, consider a three-level model for the measurement y made on occasion k for subject j within center i. Relative to the previous two-level models, an additional random center effect, denoted γ_i, is included. This term indicates the influence that the center is having on the response of the individual. We assume that the distribution of the center effects is normal with mean 0 and variance σ_γ^2. To the degree that clustering of individuals within centers influences the individual outcomes, the center effects γ_i deviate from zero, and the population variance associated with these effects σ_γ^2 increases in value. Conversely, when the clustering of individuals within centers is having little influence on the individual outcomes, the center effects γ_i will all be near zero, and the population variance σ_γ^2 will approach zero. Empirical Bayes estimation can be used for the center effects γ_i, with marginal maximum likelihood estimation of the population variance term σ_γ^2.

Treatment of the center effects as a random term in the model means that the specific centers used in the study are considered to be a representative sample from a larger population of potential centers. Conversely, if interest is only in making inferences about the specific centers of a dataset (e.g., is there a difference between center A and center B?), the center could then be regarded as a "fixed" and not a "random" effect in the model. The random-effects approach is advised when there is interest in assessing the overall effect that any potential center may have on the data and, thus, in determining the degree of variability that the center accounts for in the data. This coincides with the manner in which the center is often conceptualized, that is, the center was drawn from a population of potential centers, and the 8 or 12 centers used in the study are not the population itself.

In addition to the random center effect, center-level covariates can be included in the model to assess the influence of, say, center size on the individual responses. Interactions between center-level, individual-level, and occasion-level covariates can also be included in the model. For example, the interaction of center size (center level) by treatment group (individual level) by time (occasion level) can be included in the model to determine whether, say, treated subjects from small centers improve over time more dramatically than treated subjects from large centers. In this way, the model provides a useful method for examining and teasing out potential center-related effects.

Regarding parameter estimates, for continuous and normally distributed outcomes, Hedeker et al. (1994) noted that the fixed effects estimates are not greatly affected by the choice of

model. However, the estimates of the standard errors, which determine the significance of these parameter estimates, are influenced by the choice of model. In general, when a source of variability is present but ignored by the statistical model, the standard errors will be underestimated. Underestimation of standard errors results since the statistical model assumes that, conditional on the terms in the model, the observations are independent. However, when systematic variance is present but ignored by the model, the observations are not independent, and the amount of independent information available in parameter estimation is erroneously inflated.

Finally, when dealing with multilevel data, the number of levels of data must be considered. Often, pooling higher-order levels is determined prior to the analysis for pragmatic or conceptual reasons; at other times, the decision can be empirically tested. From the example, one could empirically argue that three-level analysis is unnecessary since the variance attributable to nesting of individuals within centers is not significant, so there is justification for pooling this additional level of the data. From a design perspective, on the other hand, there may be reason for including the center effect regardless of its statistical significance; for example, if centers were the unit of assignment in the randomization of treatment levels, one could argue that the random center term must remain in the model regardless of significance. When the variance attributable to a higher-order level is observed to be very small and nonsignificant and the sample is of moderate size, the parameter estimates and standard errors will not differ greatly whether or not the higher-order level is included in the model.

In the nonlinear case, the three-level generalization of the traditional two-level model is conceptually quite similar to that for linear mixed-effects regression models. Computationally, however, likelihood evaluation is far more complicated because the number of random effects is increased and therefore the dimensionality of the integration increases as well. Unlike the linear mixed-model, where the integrals associated with the random effect distributions do not play a role in performing maximum likelihood estimation in the marginal distribution, there is no simplification for nonlinear models such as mixed-effects binary or ordinal logistic regression, mixed-effects multinomial logistic regression, or mixed-effects Poisson regression. Here we must either evaluate the likelihood numerically (e.g., adaptive quadrature) or via simulation (Markov Chain Monte Carlo; MCMC). The interested reader is referred to Hedeker & Gibbons (2006) for detailed coverage of the technical aspects of linear and nonlinear mixed-effects regression models.

Generalized Estimating Equation Models

During the 1980s, alongside the development of mixed-effects regression models for incomplete longitudinal data, the generalized estimating equation (GEE) models were developed (Liang & Zeger 1986 and Zeger & Liang 1986). Essentially, GEE models extend generalized linear models (GLMs) to the case of correlated data. Thus, this class of models has become very popular, especially for analysis of categorical and count outcomes, although they can be used for continuous outcomes as well. One difference between GEE models and MRMs is that GEE models are based on quasi-likelihood estimation, and so the full likelihood of the data is not specified. GEE models are termed marginal models, and they model the regression of y on x and the within-subject dependency (i.e., the association parameters) separately. The term "marginal" in this context indicates that the model for the mean response depends only on the covariates of interest and not on any random effects or previous responses. In terms of missing data, GEE assumes that the missing data are missing completely at random (MCAR) as opposed to MAR, which is assumed by the models employing full-likelihood estimation.

Conceptually, GEE reproduces the marginal means of the observed data, even if some of those means have limited information because of subject dropout. Standard errors are adjusted (i.e., inflated) to accommodate the reduced amount of independent information produced by the correlation of the repeated observations over time (or within clusters). By contrast, mixed-effects models use the available data from all subjects to model temporal response patterns that would have been observed had the subjects all been measured to the end of the study. Because of this, estimated mean responses at the end of the study can be quite different for GEE versus MRM if the future observations are related to the measurements that were made during the course of the study. If the available measurements are not related to the missing measurements (e.g., following dropout), GEE and MRM will produce quite similar estimates. This is the fundamental difference between GEE and MRM; that is, the assumption that the missing data are dependent on the observed responses for a given subject during that subject's participation in the study. It is hard to imagine that a subject's responses that would have been obtained following dropout would be independent of their observed responses during the study. This leads to a preference for full-likelihood approaches over quasi- or partial-likelihood approaches, and MRM over GEE, at least for longitudinal data. There is certainly less of an argument for a preference for data that are only clustered (e.g., children nested within classrooms), in which case advantages of MAR over MCAR are more difficult to justify.

A basic feature of GEE models is that the joint distribution of a subject's response vector y_i does not need to be specified. Instead, it is only the marginal distribution of y_{ij} at each time point that needs to be specified. To clarify this further, suppose that there are two time points and suppose that we are dealing with a continuous normal outcome. GEE would only require us to assume that the distribution of y_{i1} and y_{i2} are two univariate normals, rather than assuming that y_{i1} and y_{i2} form a (joint) bivariate normal distribution. Thus, GEE avoids the need for multivariate distributions by only assuming a functional form for the marginal distribution at each time point.

METHODS TO BE AVOIDED

Unfortunately, despite progress in the use of longitudinal designs in research studies, analysis of the resulting longitudinal data has not always been commensurate with the increased value of the data. Several approaches that are of historical interest only remain in routine practice today, and methods that unnecessarily reduce the longitudinal nature of the data into a cross-sectional summary unfortunately also remain in use despite enormous statistical progress in this area.

One of the worst examples of loss of valuable statistical information is "endpoint" analysis. The idea here is that although high-quality longitudinal data may be available, if they are used at all, it is only to fill in the blanks left by patients who have discontinued the study.

Completer Analysis

In some cases a "completer analysis" is performed in which only those patients who completed the study are the focus of the analysis. Of course, in the presence of missing data produced by patients who discontinue the study, the sample that is randomized may be quite different from the sample that is analyzed. If patients who complete the study are more compliant than those who do not, valuable information regarding treatment-related effects may be obscured. Completer analyses will generally lose the critically important advantages afforded by randomization.

Last Observation Carried Forward

A second approach that remains in reasonably widespread use is to impute the measurement at the end of the study using the last available measurement obtained from the subject during the study. This approach is termed "last observation carried forward" (LOCF). Here we assume that once subjects drop out of a study, their level of response would remain unchanged. Furthermore, no distinction in the analysis is made between those patients who actually had a valid measurement at the end of the study and those subjects for whom the study endpoint was imputed based on an earlier available measurement. Although it has been suggested that the LOCF approach is conservative and that is why it continues to be used by the U.S. Food and Drug Administration for the approval of new drugs, this is clearly not the case (Lavori et al. 2008, Molenberghs et al. 2004). As noted by Lavori et al. (2008, p. 789), "The most serious effect of LOCF is that it lulls the designer into a false sense of security about the need to resolve complex issues of nonadherence, as well as the consequences of unrestrained missingness for precision and bias. This 'moral hazard' has led investigators to underdesign studies in the mistaken belief that LOCF cures all."

Despite enormous progress in the development of new and statistically more rigorous methods for analysis of longitudinal data, older and far less general methods remain in use. A historical foundation is presented by Hedeker & Gibbons (2006), who review these methods in detail.

Repeated Measures ANOVA

The traditional mixed-model ANOVA or so-called repeated-measures ANOVA was essentially a random intercept model that assumed that subjects could only deviate from the overall mean response pattern by a constant that was equivalent over time. A more reasonable view is that the subject-specific deviation is both in terms of the baseline response (i.e., intercept) and in terms of the rate of change over time (i.e., slope or set of trend parameters). This more general structure could not be accommodated by the repeated measures ANOVA. The random intercept model assumption leads to a compound-symmetric variance-covariance matrix for the repeated measurements in which the variances and covariances of the repeated measurements are constant over time. In general, we find that variances increase over time and covariances decrease as time-points become more separated in time. Finally, based on the use of least-squares estimation, the repeated measures ANOVA breaks down for unbalanced designs, such as those in which the sample size decreases over time due to subject discontinuation. Based on these limitations, the repeated measures ANOVA and related approaches should no longer be used for analysis of longitudinal data.

Multivariate Growth Curve Models

An improvement over the traditional repeated measures ANOVA was the multivariate growth curve model (Bock 1975, Potthoff & Roy 1964). The primary advantage of the multivariate analysis of variance (MANOVA) approach versus the ANOVA approach is that the MANOVA assumes a general form for the correlation of repeated measurements over time, whereas the ANOVA assumes the much more restrictive compound-symmetric form. The disadvantage of the MANOVA model is that it requires complete data. Subjects with incomplete data must be removed from the analysis, leading to potential bias. In addition, both MANOVA and ANOVA models focus on comparison of group means and provide no information regarding subject-specific growth curves. Finally, both ANOVA and MANOVA models require that the time-points are fixed across subjects (either evenly or unevenly spaced) and are treated as a classification variable in the ANOVA or MANOVA model. This precludes analysis of unbalanced designs in which different subjects are measured on different occasions. Finally, the MANOVA approach precludes use of time-varying

covariates that are often essential to modeling dynamic relationships between predictors and outcomes.

SAMPLE SIZE DETERMINATION

Despite the now widespread use of longitudinal designs in behavioral research, surprisingly little statistical research has been conducted on sample size determination for longitudinal studies. Hedeker et al. (1999) developed sample size formulas for two-level models, and more recently, Roy et al. (2007) developed a general approach to sample size determination for three-level models. All of this work is for linear mixed effects models, though methods are also emerging for nonlinear mixed models (Dang et al. 2008). Based on the work of Roy et al. (2007) and Bhaumik et al. (2009), we illustrate some of the interesting features of sample size determination for three-level mixed effects regression models using a typical psychiatric application and computed using the RMASS Web-based application (www.healthstats.org).

Bhaumik et al. (2009) consider three-level data from the now classic National Institute of Mental Health schizophrenia collaborative study, which was one of the last large-scale multicenter studies to randomize schizophrenic patients to placebo. The study consisted of nine centers and seven repeated measurements over the course of six weeks. Bhaumik et al. (2009) fitted a three-level linear mixed effects regression model to the data from two arms of the study (placebo and chlorpromazine) and used the estimated model parameters and random effect variance estimates to determine sample size for a wide variety of future studies. In the following, we summarize a few of these results of interest.

For subject-level randomization, power of 95%, attrition of 5% per week, and $ES = 0.5\ SD$ units at the end of the study, with six centers, $n = 34$ subjects per center are required for a total of 204 subjects. Decreasing power to 80% decreases the number of subjects per center to $n = 19$ or a total of 114 subjects. Note that for subject-level randomization, if we double the number of centers from 6 to 12, one-half of the number of subjects per center are required (e.g., $n = 17$ for power of 95%). This result is expected because when using subject-level randomization, the sample size formula does not involve center-level random effect variances. This is not true if we were to randomize centers to different interventions. With cluster randomization, power of 80% is achieved for a study with six centers and $n = 47$ subjects per center; a total of 282 subjects as compared to 114 subjects for subject-level randomization. Increasing the number of centers to 12 decreases the number of subjects per center to $n = 14$, or a total of only 168 subjects. Under cluster randomization, there is a substantial tradeoff between number of centers and number of subjects per center, which can affect the total number of subjects required. If we were to require power of 95%, then a minimum of eight centers are required. With eight centers, $n = 114$ subjects per center are required, or a total of 912 subjects. This is over four times as many subjects as are required for subject-level randomization ($n = 204$). However, increasing the number of centers to 12 decreases the number of subjects per center to 35 (420 total), and increasing the number of centers to 50 decreases the total sample size to $n = 250$ (i.e., $n = 5$ per center). As can be seen, cluster randomization imposes a substantial tradeoff between numbers of centers and total number of subjects.

Finally, decreasing the *ES* to *0.3 SD* units further increases sample size requirements as expected. Under subject-level randomization, power of 80% is achieved for any combination of centers and number of subjects within centers that totals to $n = 320$, and for 95% power, a total of $n = 560$ subjects are required. For cluster randomization, a minimum of 11 centers are required for power of 80% ($n = 290$ per center) and a minimum of 19 centers for power of 95% ($n = 355$ per center). More conservatively, assuming 25 centers, power of 80% is

achieved for $n = 21$ per center, and power of 95% is achieved for $n = 73$ per center. As such, cluster-level randomization increases the total number of subjects from 320 for subject-level randomization to a total of 525 subjects, assuming that 25 centers are available. For the same 25 centers, requiring power of 95% more than triples the total sample size.

Bhaumik and colleagues present the following guidelines:

> While it should be clear that sample size determination is study-specific, it is possible to provide some general guidelines. For example, consider a study involving 5 measurement occasions (e.g., baseline and 4 weekly measurements during the active treatment/intervention phase of the study), and a two-group comparison (e.g., drug versus placebo). For subject-level randomization with no variation in impact as a function of center, the total number of centers provides a negligible effect on statistical power and sample size. As such, we can compute the total number of subjects that are required for a given ES at the final time-point, under the assumption of a linear time by treatment interaction. For example, assuming subject-level randomization, five time-points, power of 80%, a Type I error rate of 5% and no attrition, to detect a one-half standard deviation unit difference at the end of the study requires a total sample size of 90 subjects or 45 per treatment arm. To detect a one-third standard deviation unit difference at the end of the study assuming all of the other conditions remained the same, a total of 200 subjects or 100 subjects per treatment arm are required (e.g., 40 subjects in each of 5 centers or 20 subjects in each of 10 centers). Adding attrition of 5% per wave increases the required sample size by about 15% and adding attrition of 10% per wave increases the required sample size by about 30% relative to no attrition.

> For cluster-level randomization, the results are more complicated. Assuming the same conditions as above, and 10 centers, to detect a one-half standard deviation unit difference at the end of the study would require a total sample size of 200 subjects (20 per center) or 100 per treatment arm. To detect a one-third standard deviation unit difference at the end of the study assuming all of the other conditions remained the same, would require a minimum of 14 centers, regardless of the number of subjects per center. Conservatively, if we increase the number of centers to 20, a total of 560 subjects (28 per center) or 280 subjects per treatment arm are required. Similar to subject-level randomization, adding attrition of 5% per wave increases the required sample size by about 15% and adding attrition of 10% per wave increases the required sample size by about 30% relative to no attrition. (Bhaumik et al. 2009, pp. 769–770)

RECENT ADVANCES IN GENERALIZED MIXED-EFFECTS REGRESSION MODELS

In the following sections, we consider several recent advances in the analysis of linear and nonlinear mixed-effects regression models. The level of statistical presentation is at a higher level than in previous sections owing to the increased complexity of these newer methodologies.

Three-Level Models

In a previous section, we provided a general overview of three-level models. In the following, we review some of the more specific statistical details corresponding to linear and nonlinear three-level models.

Sampling design—Let $\mathbf{y}_{ij}$ denote a $n_{ij} \times 1$ vector of outcomes with typical element y_{ijk}, where i denotes the level-3 units, j denotes the level-2 units nested within the i-th level-3 unit, and k denotes the level-1 units nested within $i\,j$. Assume further that there are N level-3 units so that $i = 1, 2, \ldots, N$. Within a typical level-3 unit there are n_i level-2 units, $j = 1, 2, \ldots, n_i$, and nested within $i\,j$ there are n_{ij} level-1 units, so that $k = 1, 2, \ldots, n_{ij}$. There are, therefore, $\sum_{i=1}^{N} n_i$ level-2 units and $\sum_{i=1}^{N} \sum_{j=1}^{n_i} n_{ij}$ level-1 units.

Linear Models

Let

$$y_{ijk} = \mathbf{x}'_{ijk}\boldsymbol{\beta} + \mathbf{z}'_{(3)ijk}\mathbf{v}_i + \mathbf{z}'_{(2)ijk}\mathbf{u}_{ij} + e_{ijk},$$

where $\boldsymbol{\beta}$ is an ($m \times 1$) vector of regression coefficients, and where $\mathbf{v}_i$, $\mathbf{u}_{ij}$, and e_{ijk} denote level-3, level-2, and level-1 random effects, respectively. We assume that $\mathbf{v}_1, \mathbf{v}_1, \ldots, \mathbf{v}_N$ are i.i.d. $N(\mathbf{0}, \boldsymbol{\psi})$, independent of $u_{11}, u_{12}, \ldots, u_{Nn_i}$ which are i.i.d. $N(\mathbf{0}, \boldsymbol{\Phi})$. We further assume that the $\mathbf{v}_i$ and the $\mathbf{u}_{ij}$ effects are independent of $e_{111}, e_{112}, \ldots, e_{Nn_in_{ij}}$ which are i.i.d. $N(0, \sigma^2)$.

The set of regression equations, $k = 1, 2, \ldots, n_{ij}$ can be written as

$$\mathbf{y}_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{(3)ij}\mathbf{v}_i + \mathbf{Z}_{(2)ij}\mathbf{u}_{ij} + \mathbf{e}_{ij},$$
$$j = 1, 2, \ldots, n_i,$$

where y_{ijk}, $\mathbf{x}'_{ijk}$, $\mathbf{z}'_{(3)ijk}$, $\mathbf{z}'_{(2)ijk}$ and e_{ij} are typical rows of $\mathbf{y}_{ij}$, $\mathbf{X}_{ij}$, $\mathbf{Z}_{(3)ij}$, $\mathbf{Z}_{(2)ij}$ and $\mathbf{e}_{ij}$ and where $\mathbf{X}_{ij}$, $\mathbf{Z}_{(3)ij}$, and $\mathbf{Z}_{(2)ij}$ are the design matrices for the predictors, random level-3, and random level-2 effects, respectively.

In turn, the set of regression equations is

$$\mathbf{y}_i^* = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i^*\mathbf{v}_i^* + \mathbf{e}_i^*$$

where

$$\mathbf{y}_i^* = (\mathbf{y}_{1i}, \mathbf{y}_{2i}, \ldots, \mathbf{y}'_{in_i}),$$

$$\mathbf{X}_i = \begin{pmatrix} X_{i1} \\ X_{i1} \\ \vdots \\ X_{in_i} \end{pmatrix},$$

$$\mathbf{Z}_i^* = \begin{pmatrix} \mathbf{Z}_{(3)i1} & \mathbf{Z}_{(3)i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{Z}_{(3)i1} & \mathbf{0} & \mathbf{Z}_{(2)i2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \mathbf{0} \\ \mathbf{Z}_{(3)in_i} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}_{(2)in_i} \end{pmatrix},$$

$$\mathbf{v}_i^* = (\mathbf{v}_i, \mathbf{u}_{i1}, \mathbf{u}_{i2}, \ldots, \mathbf{u}_{in_i})',$$

and

$$\mathbf{e}_i^* = (\mathbf{e}_i, \mathbf{e}_{i1}, \mathbf{e}_{i2}, \dots, \mathbf{e}_{in_i})'.$$

From the distributional assumptions imposed on $\mathbf{v}_i$, $\mathbf{u}_{ij}$ and e_{ijk}, it follows that

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

where $\boldsymbol{\mu}_i = \mathbf{X}_i \boldsymbol{\beta}$ (the fixed part of the model) and

$$\sum_i = \mathbf{Z}_i^* \mathrm{cov}(\mathbf{v}_i^*) \mathbf{Z}_i^{*'} + \sigma^2 I$$

where

$$\mathrm{cov}(\mathbf{v}_i^*) = \begin{pmatrix} \psi & 0 & \cdots & 0 \\ 0 & \Phi & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Phi \end{pmatrix}.$$

The unknown parameters are the elements of the vector $\boldsymbol{\beta}$ of fixed regression coefficient and the nonduplicated elements of the variance component matrices ψ and Φ and the level-1 error variance σ^2. The log-likelihood function of $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N$ is

$$\ln L = -\frac{1}{2} \sum_{i=1}^{N} \{ n_i \ln 2\pi + \ln | \sum_i | + tr \sum_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)' \}.$$

Nonlinear Models

A multilevel model with a nonnormal outcome variable is transformed to a linear model by using a *link* function that defines the relationship between the dependent variable η_{ijk} of the linear model and the mean μ_{ijk} of the distribution selected. More specifically, the linear model of a multilevel generalized linear model is given by

$$\eta_{ijk} = \mathbf{x}_{ijk}' \boldsymbol{\beta} + \mathbf{z}_{(2)ijk}' \mathbf{u}_{ij} + \mathbf{z}_{(3)ijk}' \mathbf{v}_i,$$

where $\mathbf{x}_{ijk}$ is a $p \times 1$ vector of predictors and $\mathbf{z}_{(2)ijk}$ is a $q \times 1$ design vector associated with the level-2 random effects $\mathbf{u}_{ij}$. Likewise, $\mathbf{z}_{(3)ijk}$ is a $r \times 1$ design vector associated with the level-3 random effects $\mathbf{v}_i$. Typically, the elements of $\mathbf{z}_{(3)ijk}$ and $\mathbf{z}_{(2)ijk}$ are subsets of the elements of $\mathbf{x}_{ijk}$.

It is further assumed that the level-3 and level-2 random effect vectors are uncorrelated and also that $\mathbf{v}_i \sim N(\mathbf{0}, \mathbf{\psi})$ and that $\mathbf{u}_{ij} \sim N(\mathbf{0}, \mathbf{\Phi})$.

Quadrature is a numeric method for evaluating multidimensional integrals. For mixed-effect models with count and categorical outcomes, the log-likelihood function is expressed as the sum of the logarithm of integrals, where the summation is over higher-level units, and the dimensionality of the integrals equals the number of random effects.

Quadrature entails the approximation of the definite integral of a function, usually stated as a weighted sum of function values at specified points within the domain of integration. Adaptive quadrature generally requires fewer points and weights to yield estimates of the model parameters and standard errors that are as accurate as would be obtained with more points and weights in nonadaptive quadrature. The reason for that is that the adaptive quadrature procedure uses the empirical Bayes means and covariances, updated at each iteration to essentially shift and scale the quadrature locations of each higher-level unit in order to place them under the peak of the corresponding integral.

A brief description of quadrature follows below, assuming a level-2 mixed effects model. Using the rules for joint and conditional distributions, it follows that

$$f(\mathbf{y}_i, \mathbf{v}_i) = f(\mathbf{y}_i | \mathbf{v}_i) f(\mathbf{v}_i),$$

and that the marginal distribution of $\mathbf{y}_i$ can be obtained as the solution to the multidimensional integral

$$f(\mathbf{y}_i) = \int_{v_1} \cdots \int_{v_r} f(\mathbf{y}_i | \mathbf{v}_i) f(\mathbf{v}) dv_1 \ldots . dv_r.$$

Since it is assumed that $\mathbf{v}_i \sim N(\mathbf{0}, \mathbf{\Phi})$, it follows, for example, that

$$f(\mathbf{v}_i) = (2\pi)^{-r/2} |\mathbf{\Phi}|^{-1/2} \exp\left[-\frac{1}{2} \mathbf{v}_i' \mathbf{\Phi}^{-1} \mathbf{v}_i \right].$$

In general, a closed-form solution to this integral does not exist. To evaluate integrals of the type described above, we use a direct implementation of Gauss-Hermite quadrature (see, e.g., Krommer & Ueberhuber 1994, section 4.2.6, and Stroud & Sechrest 1966, section 1). With this rule, an integral of the form

$$I(t) = \int f(t) \exp[-t^2] \, dt$$

is approximated by the sum

$$I(t) \approx \sum_{u=1}^{Q} w_u f(z_u),$$

where w_u and z_u are weights and nodes of the Hermite polynomial of degree Q. A Q-point adaptive quadrature rule is a quadrature rule constructed to yield an exact result for polynomials of degree $2Q - 1$, by a suitable choice of the n points x_i and n weights w_i.

Estimates of the variance components of the random effects and estimates of the fixed parameters are obtained by iteratively solving the equations

$$\frac{\partial}{\partial \gamma_k} \ln f(\mathbf{y}_i) = 0,$$

where γ_k is a typical element of the vector γ of unknown parameters $\varphi_{11}, \varphi_{21}, \ldots, \varphi_{rr}$ and β_1, $\beta_2, \ldots, \beta_p$.

Clinic	Patient	y1	y2	y3	x1	x2
1	1	0	2	1	22	−1
1	2	1	3	−9	30	1
1	3	−9	2	1	26	−1
2	1	0	1	1	23	−1
2	2	0	2	0	29	1
2	3	0	1	1	26	1
2	4	1	2	1	33	−1

To create a data set that can be analyzed within a level-3 linear model framework, dummy variables are created for each response variable in the data set. For the example above, this translates to three dummy-coded variables: $d_k = 1$ if y_k is measured, $k = 1, 2, 3$, and 0 otherwise. Using these dummy variables, we construct a new data set, shown below for clinic number 1, patients 1, 2, and 3.

Clinic	Patient	y	d1	d2	d3	x1*d1	x1*d2	x1*d3	x2*d1	x2*d2	x2*d3
1	1	0	1	0	0	22	0	0	−1	0	0
1	1	2	0	1	0	0	22	0	0	−1	0
1	1	1	0	0	1	0	0	22	0	0	−1
1	2	1	1	0	0	30	0	0	1	0	0
1	2	3	0	1	0	0	30	0	0	1	0
1	3	2	0	1	0	0	26	0	0	−1	0
1	3	1	0	0	1	0	0	26	0	0	−1

Multivariate Mixed Models

In mental health research, researchers often have data sets containing more than one response variable. A typical example is counts of inpatient (y_1), outpatient (y_2), and emergency room (y_3) visits for mental health care. There is thus a need to fit multivariate response variables to a linear mixed-effects model. It turns out that, with the use of dummy

variables, a multivariate level-2 model can be fitted to the data using a level-3 model with a single response variable and no level-1 random effects.

For the variables y_1, y_2, and y_3 considered above, the following represents a typical data set, where x_1 = *depressive severity* and x_2 = *type of insurance coverage* (coded 1 for public and −1 for private). Missing values are coded −9.

In the level-3 framework, y is the response variable, d1, d2, d3, x1*d1, …., x2*d3 are typical rows of the fixed-effects design matrix **X**. The fixed-effects part consists of intercept coefficients (corresponding to d1, d2, and d3), slope coefficients for depressive severity (corresponding to x1*d1, x1*d2, and x1*d3), and insurance coverage coefficients (corresponding to x2*d1, x2*d2, and x2*d3). Alternatively, one can use depression and insurance as level-2 covariates, in which case the data set (shown for Clinic 1, Patient 1 only) has the form:

Clinic	Patient	y	d1	d2	d3	x1	x2
1	1	0	1	0	0	22	−1
1	1	2	1	1	0	22	−1
1	1	1	1	0	1	22	−1

The difference between the two approaches is that in the first approach, different slopes are assumed for the three service utilization outcome variables, whereas we assume equal slopes for depression and equal slopes for insurance type in the second approach.

Four-Level Models

Consider a clinical study designed to measure the impact of hormone therapy on memory and cognition in elderly women. Suppose that 50 hospitals (level-4 units) participated in the study. For each of the hospitals, data are available for five types of hormone treatments (level-3 units) obtained from the female patients (level-2 units) who were tested twice a year for a period of up to six years (level-1 units).

Let y_{ijkl} denote a cognition score at occasion l for patient k on treatment j at hospital i.

A typical mixed-effects model for data of this type is

$$y_{ijkl} = \beta_0 + \beta_1 x_{1,ijkl} + \beta_2 x_{2,ijkl} + \cdots + \beta_r x_{r,ijkl} + w_i + v_{ij} + u_{ijk} + e_{ijkl},$$

where w_i denotes a level-4 (hospital-level) variance component; v_{ij}, a level-3 (treatment-level) variance component; u_{ijk}, a level-2 (patient-level) variance component; and e_{ijkl}, the level-1 measurement error. It is further assumed that there are r covariates x_1, x_2, …, x_r (such as age, weight, and percentage fat) that may influence the cognition score.

The set of regression equations can be rewritten as

$$\mathbf{y}_{ijkl} = \mathbf{X}_{ikl}\boldsymbol{\beta} + \mathbf{Z}_{(3)ikl}\mathbf{v}_i^* + \mathbf{Z}_{(2)ikl}\mathbf{v}_{ik} + \mathbf{Z}_{(1)ikl}\mathbf{e}_{ikl},$$

where

$$
\mathbf{X}_{ikl} = \begin{bmatrix} 1 & x_{1,i1kl} & \cdots & x_{r,i1kl} \\ 1 & x_{1,i2kl} & \cdots & x_{r,i2kl} \\ 1 & x_{1,i3kl} & \cdots & x_{r,i3kl} \\ 1 & x_{1,i4kl} & \cdots & x_{r,i4kl} \\ 1 & x_{1,i5kl} & \cdots & x_{r,i5kl} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix},
$$

$$
\mathbf{Z}_{(3)ikl}\mathbf{v}_i^* = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} w_i \\ v_{i1} \\ v_{i2} \\ v_{i3} \\ v_{i4} \\ v_{i5} \end{bmatrix},
$$

$$
\mathbf{Z}_{(2)ikl}\mathbf{v}_{ik} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} v_{i1k} \\ v_{i2k} \\ v_{i3k} \\ v_{i4k} \\ v_{i5k} \end{bmatrix}.
$$

We note that, except for column 1 of the design matrix $\mathbf{Z}_{(3)}$, the remaining columns correspond to dummy variables $T_1, T_2, \ldots, T_5$, where $T_j = 1$ if treatment number is j and 0 otherwise. If only treatments 2, 3, and 5 are available at hospital i, the design matrices $\mathbf{X}_{ikl}$, $\mathbf{Z}_{(3)ikl}$, and $\mathbf{Z}_{(2)ikl}$ are defined as above, but with rows 1 and 4 removed. For example,

$$
\mathbf{Z}_{(3)ikl} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.
$$

For the level-3 model to be equivalent to the level-4 model, the following patterned covariance specifications are required:

$$
\begin{bmatrix} 1 & & & & & \\ 0 & 3 & & & & \\ 0 & 0 & 3 & & & \\ 0 & 0 & 0 & 3 & & \\ 0 & 0 & 0 & 0 & 3 & \\ 0 & 0 & 0 & 0 & 0 & 3 \end{bmatrix}.
$$

The advantage of this presentation is that one can allow for cross-level correlation(s). For example, if there is reason to believe that there are differences in the way patients react to the treatments due to some hospital effect, then we may want to assume that $\operatorname{cov}(w_i, v_{ij}) \neq 0$. These covariance terms may be included in the model by using the following covariance pattern:

$$\begin{bmatrix} 1 & & & & & \\ 2 & 3 & & & & \\ 2 & 0 & 3 & & & \\ 2 & 0 & 0 & 3 & & \\ 2 & 0 & 0 & 0 & 3 & \\ 2 & 0 & 0 & 0 & 0 & 3 \end{bmatrix}.$$

Design Weights

Under the assumption that the sampling weights at a specific level are independent of the random effects at that level, Pfeffermann et al. (1998) adopted the following procedure. Consider the case of a two-level model. Denote by w_i the weight attached to the i-th level-2 unit and by $w_{j|i}$ the weight attached to the j-th level-1 unit within the i-th level-2 unit such that

$$\sum_j w_{j|i} = n_i, \quad \sum_i w_i = I$$

where I is the total number of level-2 units and $N = \Sigma_i\, n_i$ the total number of level-1 units. That is, the lower-level weights within each immediate higher-level unit are scaled to have a mean of unity, and likewise for higher levels. For each level-1 unit, we now form the final, or composite, weight

$$w_{ji} = N w_{j|i} w_i / \sum_j \sum_i w_{j|i} w_i$$
$$= N w_{j|i} w_i / \sum_i n_i w_i.$$

Denote by $\mathbf{z}_u$ and $\mathbf{z}_e$, respectively, the sets of explanatory variables defining the level-2 and level-1 random coefficients and form

$$\mathbf{z}_u^* = \mathbf{W}_i \mathbf{z}_u, \quad \mathbf{W}_i = Diag\{\omega_i^{-0.5}\}$$
$$\mathbf{z}_e^* = \mathbf{W}_{ji} \mathbf{z}_e, \quad \mathbf{W}_{ji} = Diag\{\omega_{ji}^{-0.5}\}.$$

We now carry out a standard estimation, but using $\mathbf{z}_u^*$ and $\mathbf{z}_e^*$ as the random coefficient explanatory variables. For a three-level model, with an obvious extension to notation, we have the following

$$\sum_j w_{j|ik} = n_{ik}, \quad \sum_i w_{i|k} = I_k, \quad \sum_k w_k = K,$$
$$N = \sum_i \sum_k n_{ik}, \quad I = \sum_k I_k$$
$$w_{jik} = N w_{j|ik} w_{i|k} w_k / \sum_j \sum_i \sum_k w_{j|ik} w_{i|k} w_k,$$
$$w_{ik} = I w_{i|k} w_k / \sum_i \sum_k w_{i|k} w_k.$$

Goldstein (1995) also pointed out that in survey work, analysts often have access only to the final level-1 weights w_{ji}. In this case, say for a two-level model, we can obtain the w_i by computing $w_i^{'} = W_i I / \sum_i W_i$, $W_i = (\Sigma_j w_{ji})/n_i$. For a three-level model, the procedure is carried out for each level-3 unit, and the resulting $w_{ik}^{'}$ are transformed analogously.

Examples

Subjects—The data are drawn from a natural history study of adolescent smoking. Participants included in this study were in either ninth or tenth grade at baseline and reported on a screening questionnaire 6–8 weeks prior to baseline that they had smoked at least one cigarette in their lifetimes. The majority (57.6%) had smoked at least one cigarette in the past month at baseline. A total of 461 students completed the baseline measurement wave.

The study utilized a multimethod approach to assess adolescents, including self-report questionnaires, a week-long time/event sampling method via hand-held computers (ecological momentary assessment; EMA), and detailed surveys. Adolescents carried the hand-held computers with them at all times during a data collection period of seven consecutive days and were trained both to respond to random prompts from the computers and to event record (initiate a data collection interview) in conjunction with smoking episodes. Random prompts and the self-initiated smoking records were mutually exclusive; no smoking occurred during random prompts. Questions concerned place, activity, companionship, mood, and other subjective variables. The hand-held computers dated and time-stamped each entry. Following the baseline EMA assessment period, subjects also completed similar week-long EMA assessments at 6- and 15-month follow-ups.

A question of interest concerned comparing mood from random prompts and smoking events, and the degree to which this varied was examined across the three measurement waves. For this, subjects who had at least one smoking event at two or more measurement waves were selected for analysis. In all, there were 165 such subjects with data from a total of 14,540 random prompts and smoking events; 81 of these subjects were measured at all three waves, and the remaining 84 were measured at two waves. The numbers of subjects at each wave were 152, 141, and 118, respectively. Across the waves, the average number of random prompts per subject was approximately 30 (median = 30, range = 8 to 71), 28 (median = 29, range = 4 to 47), and 28 (median = 28, range = 5 to 48), respectively. Similarly, the average number of smoking events per subject was about 6 (median = 3.5, range = 1 to 42), 5 (median = 3, range = 1 to 32), and 8 (median = 4, range = 1 to 43), respectively, across the three waves. The Spearman correlation between the number of random prompts and number of smoking events was not significant at any wave ($r = -0.16$, -0.07, and -0.05, respectively).

Dependent Measures

Negative and positive affect—Two mood outcomes were considered: measures of the subject's negative and positive affect (denoted NA and PA, respectively) at each random prompt and at each smoking episode. Both of these measures consisted of the average of several individual mood items, each rated from 1 to 10, that were identified via factor analysis. Specifically, PA consisted of the following items that reflected subjects' assessments of their positive mood just before the prompt signal: I felt happy, I felt relaxed, I felt cheerful, I felt confident, and I felt accepted by others. Similarly, NA consisted of the following items assessing pre-prompt negative mood: I felt sad, I felt stressed, I felt angry, I felt frustrated, and I felt irritable. Subjects rated each item on a 1–10 Likert-type scale, with 10 representing very high levels of the attribute. For the smoking events, participants rated their mood right after smoking. Over all prompts and events, both random and smoking, and

ignoring the clustering of the data within subjects, the mean of PA was 6.70 (sd = 1.96), 6.75 (sd = 1.92), and 6.86 (sd = 1.88) across the three waves. Similarly, for NA, the means were 3.57 (sd = 2.33), 3.43 (sd = 2.27), and 3.31 (sd = 2.18). Thus, overall, positive affect increased and negative affect decreased over time.

Inspection of the marginal distributions of these two variables indicated approximate normality for PA, but not for NA, which had a large proportion of responses with values of 1. Thus, we will describe analysis using a linear mixed-effects regression model for PA and an ordinal logistic mixed-effects regression model for NA. These models were estimated using the software program SuperMix (Hedeker et al. 2008)

Three-Level Linear Mixed-Effects Regression Model for Changes in PA Associated with Smoking Events Across Waves

Consider the following linear mixed model for the PA mood measurement y of individual i ($i = 1, 2, …, N$ level-3 subjects) at wave j ($j = 1, 2, …, n_i$ level-2 waves), and occasion k ($k = 1, 2, …, n_{ij}$ level-1 prompts and events). Let Wave denote the measurement wave (coded 0, 1, and 2.5; each unit represents a six-month time interval), and SmkE represent a variable indicating whether the occasion is from a random prompt (= 0) or a smoking event (= 1):

$$y_{ijk} = (\beta_0 + v_{0i} + u_{0ij}) + (\beta_1 + v_{1i})\text{Wave}_{ij} + (\beta_2 + v_{2i})\text{SmkE}_{ijk} + \beta_3(\text{SmkE}_{ijk} \times \text{Wave}_{ij}) + \varepsilon_{ijk}.$$

In this model, there are three random subject effects to allow subjects to vary in their intercept (v_{0i}), time trend (v_{1i}), and effect of smoking (v_{2i}) on their mood. Also, a random effect for measurement wave within subjects (u_{0ij}) is included to account for additional within-wave mood correlation of the subject responses. Note that these random effects are all deviations relative to the model fixed effects. In terms of the fixed effects, the model includes effects of wave (β_1), smoke event (β_2), and the interaction of smoke event by wave (β_3). This latter term indicates the degree to which the mood effect of smoking varies across waves. Finally, the model also includes ε_{ijk}, which is an independent (level-1) error term distributed normally with mean 0 and variance σ_ε^2.

The level-1 errors are independent conditional on the random effects (v_{0i}, v_{1i}, and v_{2i} at the subject level, and u_{0ij} at the wave level). With three random subject effects, the population distribution of intercept and slope deviations is assumed to be a trivariate normal $N(0, \Sigma_v)$, where Σ_v is the 3×3 variance-covariance matrix given as:

$$\sum_v = \begin{pmatrix} \sigma_{v_0}^2 & \sigma_{v_0 v_1} & \sigma_{v_0 v_2} \\ \sigma_{v_0 v_1} & \sigma_{v_1}^2 & \sigma_{v_1 v_2} \\ \sigma_{v_0 v_2} & \sigma_{v_1 v_2} & \sigma_{v_2}^2 \end{pmatrix}.$$

These variances indicate the heterogeneity in the population of subjects in terms of intercepts, and effects of time and smoking on mood. Likewise, the covariances represent association of these subject-specific random effects in the population of subjects. The additional level-2 variance $\sigma_{u_0}^2$ represents between-wave (and within-subjects) variation, over and above the random subject effects.

Three-Level Ordinal Logistic Mixed-Effects Regression Models for Changes in Negative Affect Associated with Smoking Events Across Waves

As indicated above, the distribution of the mood outcome for negative affect was not approximately normally distributed, and so we present an analysis treating the outcome as

an ordinal variable. Three-level models for ordinal outcomes are described in Raman & Hedeker (2005) and Liu & Hedeker (2006), building upon the development of the three-level model for binary outcomes in Gibbons & Hedeker (1997). Here, the ordinal outcome Y_{ijk} can take on values $c = 1, 2, \ldots, C$, and the model is written in terms of the cumulative logit, namely:

$$\ln\left[\frac{\Pr(Y_{ijk} \leq c)}{1 - \Pr(Y_{ijk} \leq c)}\right] = \gamma_c - [(v_{0i} + u_{0ij}) + (\beta_1 + v_{1i})\text{Wave}_{ij} + (\beta_2 + v_{2i})\text{SmkE}_{ijk} + \beta_3(\text{SmkE}_{ijk} \times \text{Wave}_{ij})]$$

with $C - 1$ strictly increasing model thresholds γ_c (i.e., $\gamma_1 < \gamma_2 \cdots < \gamma_{C-1}$). These thresholds reflect the marginal frequencies in the C categories of the ordinal outcome. For identification, either one of the thresholds or the model intercept must be fixed to zero. In the above, we have specified the latter. Otherwise, the model for NA includes the fixed and random effects as in the analysis of the continuous outcome PA. The interpretation for the covariates and random effects is in terms of the logit scale. Note that none of these carry the category subscript c, and so their effects are constant across the cumulative logits. McCullagh (1980) calls this assumption of identical odds ratios across the $C - 1$ cut-offs the proportional odds assumption.

We should note that for both PA and NA, we also investigated the decomposition of the effect of the occasion-varying SmkE on mood, as described in Begg & Parides (2003), by including the subject's mean $\overline{\text{SmkE}}_{ij}$ at each wave as an additional covariate. Notice that because SmkE_{ijk} is simply a binary variable, taking on values of 0 or 1, $\overline{\text{SmkE}}_{ij}$ simply equals the proportion of occasions (i.e., both random prompts and smoking events) at a wave that were smoking events for a subject. By so doing, the effect of SmkE indicates how a person's mood differs between a random prompt and smoking event controlling for the proportion of smoking events that the person has at the given wave. Inclusion of this additional covariate (as well as its interaction with wave) had minimal effect and did not alter the conclusions. Thus, for simplicity we leave this out in the results described below.

Results

Results for both (continuous) positive and (ordinal) negative affect are listed in Table 1.

Despite the different scales and models of these two variables, the results are relatively consistent for the two mood outcomes, albeit in opposite directions in terms of the mean. In terms of time, the effect of Wave on mood indicates that PA increases and NA decreases across the study waves. The effect of smoking on mood is such that positive mood is significantly increased and negative mood significantly decreased for smoking events, relative to random prompts. Thus, subjects' moods are significantly different, and in the more desired direction, just after smoking, relative to their random prompts. The interaction indicates that the positive mood benefit associated with smoking events decreases significantly across time, whereas the beneficial effect on NA does not vary across time. The decrease in positive mood boost over time may reflect the development of tolerance, as smoking also increases in these adolescents. The continued reduction in NA, relative to random times, following smoking may reflect relief from nicotine withdrawal, which may also be increasing as these adolescents continue to develop higher levels of nicotine dependency.

In terms of the variance parameters, there is considerable subject heterogeneity for both PA and NA in terms of their intercepts and slopes for Wave and SmkE. Thus, the effects of time and smoking on mood vary considerably across subjects. Also, there is significant between-

wave mood variation, both for PA and NA, in addition to the variation at the subject level. The covariances are in the same direction for PA and NA, albeit none of the covariances for the ordinal NA reach statistical significance. The negative covariance of the intercept and SmkE effect is particularly interesting. For PA, it suggests that subjects with higher random PA have less PA benefit (i.e., lower PA scores) associated with a smoking event. For NA, this negative covariance suggests that subjects with higher random NA have greater NA benefit (i.e., lower NA scores) for smoking events. Although these could be due to ceiling and floor effects of measurement, respectively, these negative covariances are consistent with previously reported results (Hedeker et al. 2009).

EXAMPLE APPLICATIONS IN THE BEHAVIORAL SCIENCES

Applications of mixed-effects regression models are steadily increasing and can be found in many different fields, including studies on alcohol (Curran et al. 1997), smoking (Niaura et al. 2002), HIV/AIDS (Gallagher et al. 1997), drug abuse (Carroll et al. 1994, Halikas et al. 1997), psychiatry (Elkin et al. 1995, Serretti et al. 2000), and child development (Campbell & Hedeker 2001, Huttenlocher et al. 1991), to name a few. Not only do these articles illustrate the wide applicability of mixed-effects regression models, they also give a sense of how mixed-effects regression model results are typically reported in the various literatures. Thus, they can be very useful for investigators who are new to mixed-effects regression models and their usage.

SUMMARY

In summary, generalized mixed-effects regression models have fundamentally changed the way we conceptualize the analysis of longitudinal data. The ability to model and compare longitudinal response patterns for continuous and categorical outcomes using a unified family of statistical models and to be able to estimate person-specific trends are major strengths of this general methodology. Robustness to missing data both under default MAR assumptions and extensions for NMAR further enhance the utility of these full likelihood procedures. Current limitations are largely related to the number of random effects that can be included in nonlinear models, where an explicit representation of the high dimensional integral must be obtained. Full Bayesian approaches provide one possible solution to this problem, and this should be an area of further statistical research.

Generalized mixed-effects regression models extend beyond analysis of longitudinal data and have been widely used for analysis of non-longitudinal clustered data as well. More recently, these models have even been applied to large-scale spontaneous reporting system data as tools to evaluate the safety of pharmaceuticals (Gibbons et al. 2008).

SUMMARY POINTS

1. Generalized mixed-effects regression models (MRMs) are ideally suited to analysis of longitudinal data.

2. MRMs are available for continuous, binary, ordinal, nominal, count, and time-to-event data.

3. Linear MRMs are less computationally intensive than nonlinear MRMs because the integrals corresponding to the random effects are not a part of the estimation.

4. Full likelihood methods such as MRM are more robust to the presence of missing data than are partial likelihood methods such as generalized estimating equations.

5. Extensions of the general MRM include three and four levels of clustering, multivariate extensions, and the inclusion of design weights.

6. New computer software such as SuperMix is now available for routine estimation of two-level and three-level linear (continuous normal) and nonlinear (binary, ordinal, nominal, count, and time-to-event) models.

Acknowledgments

Glossary

LOCF	last observation carried forward
MAR	missing at random
NMAR	not missing at random
GEE	generalized estimating equations
ANOVA	analysis of variance
MRMs	mixed-effects regression models
MCMC	Markov Chain Monte Carlo
GLMs	generalized linear models
MCAR	missing completely at random
MANOVA	multivariate analysis of variance
NA	negative affect
PA	positive affect

LITERATURE CITED

Begg MB, Parides MK. Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. Stat Med. 2003; 22:2591–602. [PubMed: 12898546]

Bhaumik DK, Roy A, Aryal S, Hur K, Duan N, et al. Sample size determination for studies with repeated continuous outcomes. Psychiatr Ann. 2009; 38:765–71. [PubMed: 19756212]

Bock, RD. Multivariate Statistical Methods in Behavioral Research. New York: McGraw-Hill; 1975.

Bock, RD. The discrete Bayesian. In: Wainer, H.; Messick, S., editors. Modern Advances in Psychometric Research. Hillsdale, NJ: Erlbaum; 1983a. p. 103-15.

Bock, RD. Within-subject experimentation in psychiatric research. In: Gibbons, RD.; Dysken, MW., editors. Statistical and Methodological Advances in Psychiatric Research. New York: Spectrum; 1983b. p. 59-90.

Bock, RD. Measurement of human variation: a two-stage model. In: Bock, RD., editor. Multilevel Analysis of Educational Data. New York: Academic; 1989. p. 319-42.

Bryk, AS.; Raudenbush, SW. Hierarchical Linear Models: Applications and Data Analysis Methods. Newbury Park, CA: Sage; 1992.

Campbell SK, Hedeker D. Validity of the test of infant motor performance for discriminating among infants with varying risks for poor motor outcome. J Pediatr. 2001; 139:546–51. [PubMed: 11598602]

Carroll KM, Rounsaville BJ, Nich C, Gordon LT, Wirtz PW, Gawin F. One-year follow-up of psychotherapy and pharmacotherapy for cocaine. Depend Arch Gen Psychiatry. 1994; 51:989–97.

Chi EM, Reinsel GC. Models for longitudinal data with random effects and AR(1) errors. J Am Stat Soc. 1989; 84:452–59.

Conaway MR. Analysis of repeated categorical measurements with conditional likelihood methods. J Am Stat Assoc. 1989; 84:53–61.

Curran PJ, Stice E, Chassin L. The relation between adolescent and peer alcohol use: a longitudinal random coefficients model. J Consult Clin Psychol. 1997; 65:130–40. [PubMed: 9103742]

Dang Q, Mazumbar S, Houck PR. Sample size and power calculations based on generalized linear mixed models with correlated binary outcomes. Comput Methods Programs Biomed. 2008; 91:122–27. [PubMed: 18462826]

de Leeuw J, Kreft I. Random coefficient models for multilevel analysis. J Educ Stat. 1986; 11:57–85.

Dempster AP, Rubin DB, Tsutakawa RK. Estimation in covariance component models. J Am Stat Soc. 1981; 76:341–53.

Diggle, PJ.; Heagerty, P.; Liang, K-Y.; Zeger, SL. Analysis of Longitudinal Data. 2. New York: Oxford Univ. Press; 2002.

Elkin I, Gibbons RD, Shea MT, Sotsky SM, Watkins JT, et al. Initial severity and differential treatment outcome in the NIMH Treatment of Depression Collaborative Research Program. J Consult Clin Psychol. 1995; 63:841–47. [PubMed: 7593878]

Fitzmaurice, GM.; Laird, NM.; Ware, JH. Applied Longitudinal Analysis. New York: Wiley; 2004.

Gallagher TJ, Cottler LB, Compton WM, Spitznagel E. Changes in HIV/AIDS risk behaviors in drug users in St. Louis: applications of random regression models. J Drug Issues. 1997; 27:399–416.

Gibbons, RD. PhD thesis. Univ. Chicago; 1981. Trend in correlated proportions.

Gibbons RD, Bock RD. Trend in correlated proportions. Psychometrika. 1987; 52:113–24.

Gibbons RD, Hedeker D. Random-effects probit and logistic regression models for three-level data. Biometrics. 1997; 53:1527–37. [PubMed: 9423267]

Gibbons RD, Hedeker D, Waternaux CM, Davis JM. Random regression models: a comprehensive approach to the analysis of longitudinal psychiatric data. Psychopharmacol Bull. 1988; 24:438–43. [PubMed: 3153505]

Gibbons RD, Segawa E, Karabatsos G, Amatya AK, Bhaumik DK, et al. Random-effect Poisson regression analysis of adverse event reports: the relationship between antidepressants and suicide. Stat Med. 2008; 27:1814–33. [PubMed: 18404622]

Goldstein H. Nonlinear multilevel models, with an application to discrete response data. Biometrika. 1991; 78:45–51.

Goldstein, H. Multilevel Statistical Models. 2. New York: Halstead Press; 1995.

Halikas JA, Crosby RD, Pearson VL, Graves NM. A randomized double-blind study of carbamazepine in the treatment of cocaine abuse. Clin Pharmacol Ther. 1997; 62:89–105. [PubMed: 9246023]

Hedeker, D. PhD thesis. Univ. Chicago; 1989. Random regression models with autocorrelated errors.

Hedeker, D.; Gibbons, RD. Longitudinal Data Analysis. New York: Wiley; 2006.

Hedeker, D.; Gibbons, RD.; Du Toit, SHC.; Patterson, D. SuperMix—a program for mixed-effects regression models. Chicago: Sci. Software Int; 2008.

Hedeker D, Gibbons RD, Flay BR. Random-effects regression models for clustered data: with an example from smoking prevention research. J Consult Clin Psychol. 1994; 62:757–65. [PubMed: 7962879]

Hedeker D, Gibbons RD, Waternaux C. Sample size estimation for longitudinal designs with attrition. J Educ Behav Stat. 1999; 24:70–93.

Hedeker D, Mermelstein RJ, Berbaum ML, Campbell RT. Modeling mood variation associated with smoking: an application of a heterogeneous mixed-effects model for analysis of ecological momentary assessment (EMA) data. Addiction. 2009; 104:297–307. [PubMed: 19149827]

Hui SL, Berger JO. Empirical Bayes estimation of rates in longitudinal studies. J Am Stat Assoc. 1983; 78:753–59.

Huttenlocher JE, Haight W, Bryk AS, Seltzer M. Early vocabulary growth: relation to language input and gender. Dev Psychol. 1991; 27:236–48.

Krommer, AR.; Ueberhuber, CW. Numerical Integration on Advanced Computer Systems. New York: Springer-Verlag; 1994.

Laird NM. Missing data in longitudinal studies. Stat Med. 1988; 7:305–15. [PubMed: 3353609]

Laird NM, Ware JH. Random effects models for longitudinal data. Biometrics. 1982; 38:963–74. [PubMed: 7168798]

Lavori PW, Brown CH, Duan N, Gibbons RD, Greenhouse J. Missing data in longitudinal clinical trials—Part A: design and conceptual issues. Psychiatr Ann. 2008; 38:784–92. [PubMed: 19756211]

Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73:13–22.

Little RJA. Modeling the drop-out mechanism in repeated-measures studies. J Am Stat Assoc. 1995; 90:1112–21.

Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. 2. New York: Wiley; 2002.

Liu LC, Hedeker D. A mixed-effects regression model for longitudinal multivariate ordinal data. Biometrics. 2006; 62:261–68. [PubMed: 16542254]

Longford NT. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. Biometrika. 1987; 74:817–27.

Longford, NT. Random Coefficient Models. New York: Oxford Univ. Press; 1993.

McCullagh P. Regression models for ordinal data (with discussion). J R Stat Soc Ser B. 1980; 42:109–42.

Molenberghs GM, Thijs H, Jansen I, Beunckens C, Kenward MG, et al. Analyzing incomplete longitudinal clinical trial data. Biostatistics. 2004; 5:445–64. [PubMed: 15208205]

Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. Int Stat Rev. 1991; 59:25–35.

Niaura R, Spring B, Borrelli B, Hedeker D, Goldstein MG, et al. Multicenter trial of fluoxetine as an adjunct to behavioral smoking cessation treatment. J Consult Clin Psychol. 2002; 70:887–96. [PubMed: 12182272]

Pfeffermann D, Skinner CJ, Goldstein H, Holmes DJ, Rasbash J. Weighting for unequal selection probabilities in multilevel models (with discussion). J R Stat Soc. 1998; 60B:23–40.

Potthoff RF, Roy SN. A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika. 1964; 51:313–16.

Raman R, Hedeker D. Mixed-effects regression models for three-level ordinal response data. Stat Med. 2005; 24:3331–45. [PubMed: 16206246]

Raudenbush, SW.; Bryk, AS. Hierarchical Linear Models. 2. Thousand Oaks, CA: Sage; 2002.

Roy A, Bhaumik DK, Aryal S, Gibbons RD. Sample size determination for hierarchical longitudinal designs with differential attrition rates. Biometrics. 2007; 63:699–707. [PubMed: 17825003]

Rubin DB. Inference and missing data. Biometrika. 1976; 63:581–92.

Serretti A, Lattuada E, Zanardi R, Franchini L, Smeraldi E. Patterns of symptom improvement during antidepressant treatment of delusional depression. Psychiatry Res. 2000; 94:185–90. [PubMed: 10808044]

Singer, JD.; Willett, JB. Applied Longitudinal Data Analysis. New York: Oxford Univ. Press; 2003.

Stiratelli R, Laird NM, Ware JH. Random-effects models for serial observations with binary response. Biometrics. 1984; 40:961–71. [PubMed: 6534418]

Strenio JF, Weisberg HI, Bryk AS. Empirical Bayes estimation of individual growth curve parameters and their relationship to covariates. Biometrics. 1983; 39:71–86. [PubMed: 6871364]

Stroud, AH.; Sechrest, D. Gaussian Quadrature Formulas. Englewood Cliffs, NJ: Prentice Hall; 1966.

Verbeke, G.; Molenberghs, G. Linear Mixed Models for Longitudinal Data. New York: Springer; 2000.

Wolfinger RD. Covariance structure selection in general mixed models. Commun Stat Simulation Comput. 1993; 22:1079–106.

Wong GY, Mason WM. The hierarchical logistic regression model for multilevel analysis. J Am Stat Assoc. 1985; 80:513–24.

Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986; 42:121–30. [PubMed: 3719049]

Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. Biometrics. 1988; 44:1049–60. [PubMed: 3233245]
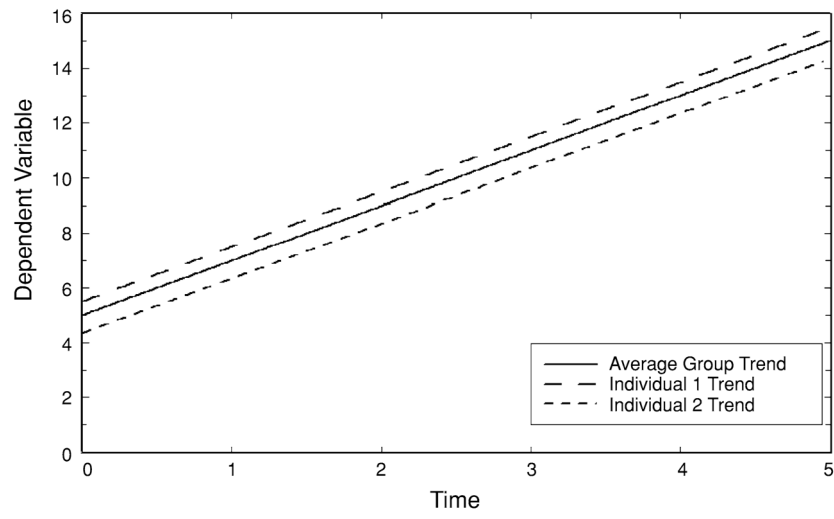
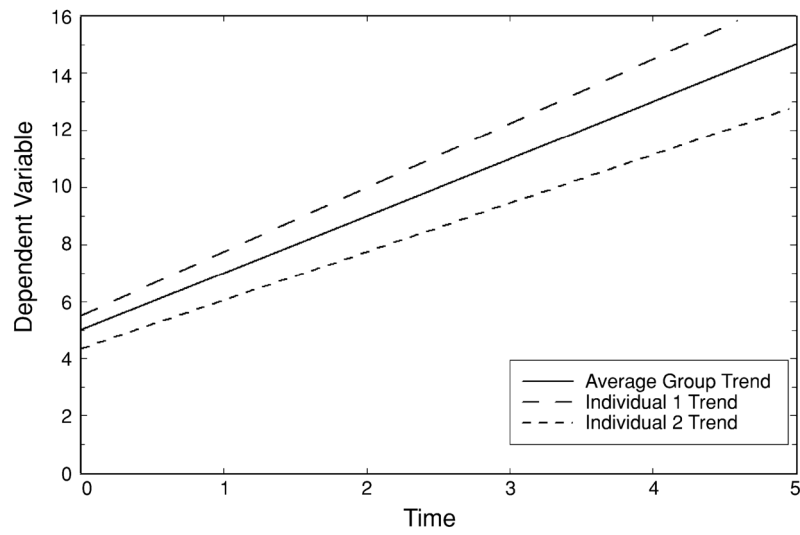**Figure 1.**
Random intercept model.
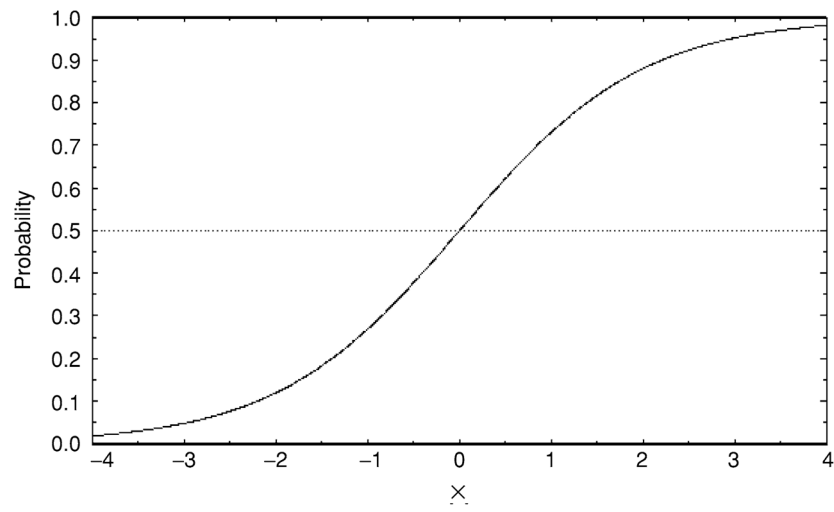
Figure 2.
Random intercept and trend model.

Figure 3.
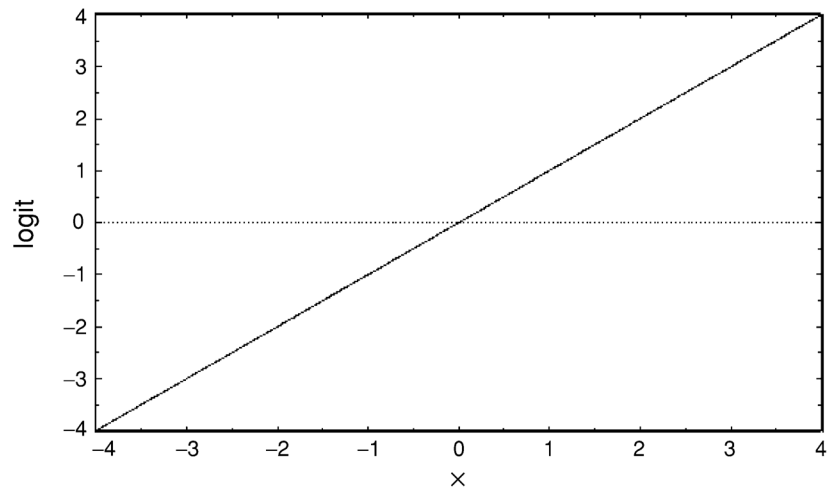Logistic relationship between *x* and the response probability.

Figure 4.
Linear relationship between *x* and the logit.

Table 1

Positive and negative affect model estimates, standard errors (se), and *p*-values

Parameter	Positive affect linear mixed model			Negative affect ordinal logistic mixed model		
	Estimate	se	*p* <	Estimate	se	*p* <
Intercept β_0	6.570	0.093	0.0001			
$Wave_{ij}$ β_1	0.091	0.041	0.03	−0.092	0.037	0.02
$SmkE_{ijk}$ β_2	0.454	0.061	0.0001	−0.320	0.077	0.0001
$SmkE_{ijk} \times Wave_{ij}$ β_3	−0.106	0.035	0.003	0.035	0.044	0.43
Intercept variance (level-3) $\sigma^2_{v_0}$	1.157	0.168	0.0001	2.603	0.374	0.0001
Wave variance (level-3) $\sigma^2_{v_1}$	0.089	0.032	0.006	0.167	0.062	0.007
SmkE variance (level-3) $\sigma^2_{v_2}$	0.127	0.037	0.001	0.198	0.073	0.007
Intercept, Wave covariance $\sigma_{v_0v_1}$	−0.043	0.054	0.44	−0.183	0.117	0.12
Intercept, SmkE covariance $\sigma_{v_0v_2}$	−0.169	0.059	0.005	−0.215	0.112	0.06
Wave, SmkE covariance $\sigma_{v_1v_1}$	−0.051	0.269	0.03	−0.062	0.043	0.14
Between-wave (level-2) variance $\sigma^2_{u_0}$	0.284	0.047	0.0001	0.554	0.094	0.0001
Error (level-1) variance σ^2_{ε}	2.230	0.027	0.0001			