Tutorial in Biostatistics

Statistics in Medicine

Received 28 July 2008, Accepted 30 September 2009 Published online 10 December 2009 in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3775

Real longitudinal data analysis for real people: Building a good enough mixed model

Jing Cheng,[a]*† Lloyd J. Edwards,[b] Mildred M. Maldonado-Molina,[c] Kelli A. Komro[c] and Keith E. Muller[a]

Mixed effects models have become very popular, especially for the analysis of longitudinal data. One challenge is how to build a *good enough* mixed effects model. In this paper, we suggest a systematic strategy for addressing this challenge and introduce easily implemented practical advice to build mixed effects models. A general discussion of the scientific strategies motivates the recommended five-step procedure for model fitting. The need to model both the mean structure (the fixed effects) and the covariance structure (the random effects and residual error) creates the fundamental flexibility and complexity. Some very practical recommendations help to conquer the complexity. Centering, scaling, and full-rank coding of all the predictor variables radically improve the chances of convergence, computing speed, and numerical accuracy. Applying computational and assumption diagnostics from univariate linear models to mixed model data greatly helps to detect and solve the related computational problems. Applying computational and assumption diagnostics from the univariate linear models to the mixed model data can radically improve the chances of convergence, computing speed, and numerical accuracy. The approach helps to fit more general covariance models, a crucial step in selecting a credible covariance model needed for defensible inference. A detailed demonstration of the recommended strategy is based on data from a published study of a randomized trial of a multicomponent intervention to prevent young adolescents' alcohol use. The discussion highlights a need for additional covariance and inference tools for mixed models. The discussion also highlights the need for improving how scientists and statisticians teach and review the process of finding a good enough mixed model. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: longitudinal data analysis; mixed effects models; model building

1. Introduction

Longitudinal studies, in which outcomes are measured repeatedly over time on the same subject, are widely conducted in various applications and often provide better insights into the processes of interest than the cross-sectional studies. Longitudinal studies allow addressing questions about changes over time in the response and changes in the relationship of the response to subjects' characteristics. Mixed effects models [1, 2] have become very popular for answering such questions.

In practice, when conducting analysis for longitudinal data with mixed models, investigators need to handle a variety of modeling tasks. Most often, investigators have many potential predictors and would like to build the 'best' mixed model. In turn, problems often emerge that must be addressed when the model does not fit. Similarly, the investigators may want to judge the validity or reliability of a model. In practice, we agree with Box and Draper's comment [3]) that 'Essentially, all models are wrong, but some are useful.' Therefore, we choose to refer to one common underlying question in the longitudinal data analysis: how do we build a *good enough* mixed model?

[a] *Division of Biostatistics, Department of Epidemiology and Health Policy Research, University of Florida College of Medicine, FL, U.S.A.*
[b] *Department of Biostatistics, School of Public Health, The University of North Carolina at Chapel Hill, NC, U.S.A.*
[c] *Department of Epidemiology and Health Policy Research, and Institute for Child Health Policy, University of Florida College of Medicine, FL, U.S.A.*
* *Correspondence to: Jing Cheng, 1329 SW 16th Street, Room 5130, Gainesville, FL 32610, U.S.A.*
† *E-mail: jcheng@biostat.ufl.edu*

Contract/grant sponsor: NIH/NIA; contract/grant numbers: P30-AG028740, R01-AG026556-01A2
Contract/grant sponsor: NIH/NIDCR; contract/grant numbers: US4-DE019261, U54-DE019261
Contract/grant sponsor: NIH/NCRR; contract/grant number: K30-RR022258
Contract/grant sponsor: NIH/NHLBI; contract/grant number: R01-HL091005
Contract/grant sponsor: NIH/NIAAA; contract/grant numbers: R01 AA013458, R01 AA016549, K01 AA017480

We use the description 'good enough' to distinguish it from the 'best' model, which in turn should be distinguished from the 'correct' model. Chatfield [4] and Draper [5] provided excellent and insightful discussions on the wider framework of finding the best and correct models. Example concerns include the following. The correct model in the population may be nonlinear, while the modeling process based on the sample data at hand may only accommodate linear functions. In contrast, a good enough linear model (including, for example, polynomial terms) may adequately approximate the correct model. The correct model for the entire population may be more complex than needed for the subpopulation from which the sample has been drawn. In contrast, a good enough model must be drawn from only those models *estimable with the data in hand*. Among the class of estimable models, one may be the best. In contrast, there may be many models that are close enough to the best model to be good enough for the purposes of the analysis at hand.

Finding a good enough model follows good statistical practice and does not allow or encourage poor data analysis. Most importantly, a good enough model must adequately satisfy all of the assumptions underlying the model-fitting process. A good enough model does not waste any substantial amount of information available in the data. At the same time, a good enough model does not extrapolate too far beyond the information available. Finally, a good enough model must provide defensible inference with credible accuracy.

In the univariate and multivariate contexts, some strategies have been proposed to select a good enough model [6–9]. For longitudinal data, Verbeke and Molenberghs [10] and Littell *et al.* [11] introduced some guidelines on building mixed models. In the present paper, we suggest a systematic framework for building a good enough mixed model for the longitudinal data in practice, and then illustrate the strategy with analysis of real data. Note that we do not review analytic approaches for the longitudinal studies. For that information, please refer to Verbeke and Molenberghs [10], Raudenbush [12], Collins [13], and Davidian [14]. Instead, we focus on mixed effects models and give both an overall strategic framework and a practical advice for fitting these models.

The remainder of the paper is organized as follows. In Section 2, we briefly introduce mixed models. In Section 3, we give a systematic framework for building a mixed effect model. In Section 4, we provide our practical advice for fitting mixed models so as to either avoid or diagnose and correct problems. We illustrate the process in Section 5 with a summary of results from a published study of alcohol use among the adolescents. Finally, we close with conclusions and a reprise of open questions.

2. Mixed effects models

2.1. The linear mixed model with the Gaussian errors

Using the notation of Muller and Stewart [8], with N independent sampling units (often *persons* in practice), the linear mixed model for person i may be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{d}_i + \boldsymbol{e}_i \tag{1}$$

where $\boldsymbol{y}_i$ is a $p_i \times 1$ vector of observations on person i at the p_i follow-up time points (the baseline measurement is used as a covariate), $\boldsymbol{X}_i$ is a $p_i \times q$ known, constant design matrix for person i with full column rank q, whereas $\boldsymbol{\beta}$ is a $q \times 1$ vector of unknown, constant, population parameters (fixed effects, the same for all subjects). In addition to the fixed effects as we see in the ordinary linear models, the linear mixed effects model also includes subject-specific random effects. Also $\boldsymbol{Z}_i$ is a $p_i \times m$ known, constant design matrix with rank m for person i corresponding to the $m \times 1$ vector of unknown random effects $\boldsymbol{d}_i$, whereas $\boldsymbol{e}_i$ is a $p_i \times 1$ vector of unknown random errors. Gaussian $\boldsymbol{d}_i$ and $\boldsymbol{e}_i$ are independent with mean $\boldsymbol{0}$ and

$$\mathscr{V}\left(\begin{bmatrix} \boldsymbol{d}_i \\ \boldsymbol{e}_i \end{bmatrix}\right) = \begin{bmatrix} \Sigma_{\boldsymbol{d}_i}(\tau_{\boldsymbol{d}}) & \boldsymbol{0} \\ \boldsymbol{0} & \Sigma_{\boldsymbol{d}_i}(\tau_{\boldsymbol{e}}) \end{bmatrix}$$

where $\mathscr{V}(\cdot)$ is the covariance operator, whereas $\Sigma_{\boldsymbol{d}_i}(\tau_{\boldsymbol{d}})$ and $\Sigma_{\boldsymbol{e}_i}(\tau_{\boldsymbol{e}})$ are positive-definite, symmetric $m \times m$ and $p_i \times p_i$ covariance matrices, respectively. Therefore the linear mixed model (1) implies that the marginal distribution of $\boldsymbol{y}_i$ is normal with mean $\boldsymbol{X}_i\boldsymbol{\beta}$ and variance $\Sigma_i = \boldsymbol{Z}_i \Sigma_{\boldsymbol{d}_i}(\tau_{\boldsymbol{d}}) \boldsymbol{Z}_i' + \Sigma_{\boldsymbol{e}_i}(\tau_{\boldsymbol{e}})$, and the conditional distribution of $\boldsymbol{y}_i | \boldsymbol{d}_i$ is normal with mean $\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{d}_i$ and variance $\Sigma_{\boldsymbol{e}_i}(\tau_{\boldsymbol{e}})$. We assume that Σ_i can be characterized by a finite set of parameters represented by an $r \times 1$ vector τ, which consists of the unique parameters in $\tau_{\boldsymbol{d}}$ and $\tau_{\boldsymbol{e}}$.

2.2. The generalized linear mixed model

We describe a generalized linear mixed model (GLMM) with a (multivariate) normal mixing distribution for the random effects. As noted by Tuerlinckx *et al.* [15], this is the model most often applied in practice. For more details, see Breslow and Clayton [16] and Tuerlinckx *et al.* [15]. With N independent sampling units and conditionally on the random effects $\boldsymbol{d}_i (m \times 1)$, assume that the responses y_{ij} of $\boldsymbol{y}_i$ are independent with density function that is a member of the exponential family, i.e.

$$f(y_{ij} | \boldsymbol{d}_i) = \exp\{[y_{ij}\theta_{ij} - b(\theta_{ij})]/a(\phi) + c(y_{ij}, \phi)\}$$

for some functions a, b, and c. The conditional mean is $E(y_{ij}|\boldsymbol{d}_i)=b'(\theta_{ij})$ and conditional variance is $\mathscr{V}(y_{ij}|\boldsymbol{d}_i)=b''(\theta_{ij})a(\phi)$. Unlike the linear mixed model, the GLMM assumes a linear regression model on the conditional mean by a function

$$g[E(y_{ij}|\boldsymbol{d}_i)]=\boldsymbol{x}'_{ij}\boldsymbol{\beta}+\boldsymbol{z}'_{ij}\boldsymbol{d}_i \tag{2}$$

where $g(\cdot)$ is referred to as a link function, e.g. logit link for the binomial data and log link for the count data. $\boldsymbol{\beta}$ is a $q\times 1$ vector of unknown fixed effect parameters, $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are $q\times 1$ and $m\times 1$ vectors of fixed and random effect explanatory variables. The $m\times 1$ subject-specific random effects $\boldsymbol{d}_i$ are assumed to be sampled from a multivariate normal distribution with mean $\mathbf{0}$ and $m\times m$ covariance matrix $\Sigma_{\boldsymbol{d}_i}(\tau_{\boldsymbol{d}})$ that depends on a $(r\times 1)$ vector $\tau_{\boldsymbol{d}}$ of unknown variance components. The conditional variance can be written as $\mathscr{V}(y_{ij}|\boldsymbol{d}_i)=\phi a_{ij}\mathscr{V}[E(y_{ij}|\boldsymbol{d}_i)]$, where a_{ij} is a known constant.

2.3. Rationale underlying mixed models

An ordinary linear model analysis assumes independence between any two observations. However, two appealing design features, repeated measures and cluster sampling, both create correlations among some observations and therefore require a more general model. In longitudinal data (repeated measures over time), the assumption of independence between two observations from different subjects typically remains valid when subjects are randomly and individually selected from populations. In contrast, two observations from the same subject will typically be correlated by sharing the same characteristics and therefore are not independent. In purely longitudinal studies, the subject defines the independent sampling unit and the measurement of a particular subject at a particular time defines the observational unit.

Cluster sampling often provides the best design for community-based and genetic research. A study with subjects from many families and two or more subjects per family provides a common and important example in which the family defines the independent sampling unit, not the subject. In turn, with one measurement per subject, the subject defines the observational unit. In a more complex design with each family member measured two or more times, the measurement at a single time for a single person defines the observational unit, whereas the family again defines the independent sampling unit. Furthermore, in such designs, we have observations from the same subject correlated and subjects from the same family correlated, where subject and family are two levels of nested clusters. In Sections 3 and 4, we will focus on dealing with correlation due to one level of clustering. In Section 5, we will discuss on handling the correlation due to nested clusters with a real example.

Multivariate or mixed effects models are needed to account for the correlations among groups of observations. Mixed effects models prove especially helpful for such data because they allow accounting for and taking advantage of the structured patterns of correlation that such designs induce. In contrast to the multivariate models, they also provide convenient modeling of some types of missing or mis-timed data, repeated covariates, and heterogeneity between groups. Multivariate models have advantages, such as ease of use, numerical stability, and accuracy of inference (confidence intervals and hypothesis tests), at least in small samples.

Mixed effects models typically include both fixed (mean model) and random (covariance model) effects and have become a primary method for longitudinal data analysis in order to account for the within-subject or within-cluster correlations. Popular models include linear mixed effects models (1) for normally distributed data, GLMMs (2) for non-normal data (e.g. binomial, poisson, etc., exponential family), and nonlinear mixed effects models for more general individual trajectories.

The coefficients for fixed effects ($\boldsymbol{\beta}$) play the same role as the coefficients in ordinary univariate regression models. They provide estimates of the average response in a group, which is, therefore, a population-specific estimate (i.e. the same for all subjects in the group), and is usually the primary interest for clinical trials and much epidemiologic and health services research. The fixed effects can be most simply understood as defining the means for a population and sub-groups defined by categorical variables. With continuous predictors, such as age and weight, the fixed effects define the expected value (regression) function. The methods and illustrations in the present paper all concentrate on the fixed effects.

Unlike ordinary regression models, in addition to the fixed effects, a mixed effects model uses subject-specific random effects $\boldsymbol{d}_i$ to account for variance heterogeneity among responses from different individuals as well as covariance pattern among responses within an individual. The assumption of random selection of individuals from larger populations leads to describing the subject-specific effects as random effects. Furthermore, the subject-specific random effects are assumed to follow some distribution with a covariance structure to account for the within-subject correlation. In some types of genetic research, especially in breeding experiments and other evaluations of the relative importance of sources of variation, the random effects in the covariance model take center stage. However, even when the covariance structure is not the primary interest, an appropriate covariance model is essential to obtain valid inference for the fixed effects.

Mixed effects models can be viewed as a result of a two-stage analysis. For example, for a linear mixed effects model, in the first stage, an appropriate function of predictors approximates each observed longitudinal profile allowing subject-specific regression coefficients:

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

where $\boldsymbol{\beta}_i$ is a $q\times 1$ vector of unknown subject-specific regression coefficients. In the second stage, another model explains the observed variability between the subjects:

$$\boldsymbol{\beta}_i=\boldsymbol{K}_i\boldsymbol{\beta}+\boldsymbol{d}_i$$

where $\boldsymbol{K}_i$ is a $q\times p$ matrix of known covariates. Note that $\boldsymbol{X}_i=\boldsymbol{Z}_i\boldsymbol{K}_i$ when we combine the two-stage models. When most variability among the measurements reflects between-subject variability, the two-stage approach helps to construct an appropriate mixed effects model. In turn, when variability depends on predictors that change within-subject, such as time, a valid mixed effects model requires an appropriate covariance structure for the residual in addition to the covariance structure modeled through the random effects.

3. Overall scientific strategies

3.1. Overview of strategies

In this section, we describe the framework of a systematic process for building a mixed effects model. The overall strategies are the same as those for the univariate models with the Gaussian errors described by Muller and Fetterman [7]. One concern in the process of model building is the type I error inflation from testing multiple models. We recommend using either one of two approaches.

- Conduct separate exploratory and confirmatory analyses by using truly independent multiple samples. Having data from two separate studies provides the best basis for the approach. Alternately, the data from a single study may be randomly split into two parts, one for exploratory analysis and another for confirmatory analysis.
- Completely specify limited tests and a testing sequence, and control type I error by multiple comparison procedures.

Regardless of which of the two overall approaches is chosen, we recommend five steps for the model selection process.

1. Specify the maximum model to be considered.
2. Specify a criterion of goodness of fit of a model.
3. Specify a predictor selection strategy.
4. Conduct the analysis.
5. Evaluate the reliability of the model chosen.

Note that the five steps describe the general strategies for model building and should be applied to linear and nonlinear models, including widely used univariate, multivariate, and mixed effects models, for any type of response distribution. Muller and Fetterman [7] provided a detailed exposition of the entire process in the context of univariate linear models with the Gaussian errors.

Any mixed effects model actually requires choosing two distinct models, one for the means and the other for the covariances. We focus almost exclusively on selecting models for the fixed effects, and hence the means of the response values. However, most of the principles apply directly to building models of the covariance matrix as well. The dependence of accurate inference for fixed effects on a valid covariance model implies an ideal approach of starting from the maximum model for fixed and random effects.

3.2. Step 0: State the scientific goal and model space

Every scientific study starts with a question that implies a goal of finding new knowledge. In the present paper, we focus on transforming the question ‘What is the best model?’ to ‘What is a good enough model?’ The best model comes from some space of possible models, perhaps ill-defined, perhaps well-defined. Starting from the scientific goal, jumping completely over the processes of study design and data collection brings us to a particular set of data and a specific question. A good enough model comes from the space of estimable models for the data at hand, which define and limit the possible models and paths to the scientific goal.

Despite the limitations of the data at hand, in some cases an inadequate or incorrect model can often still be diagnosed. In the subsequent sections we indirectly allude to a variety of assumption diagnostics such as tests for non-independence, normality, homogeneity, nonlinearity, and goodness of fit. Such diagnostics can allow, but not guarantee, discovering that the maximum model we start with does not contain the correct model, even when the correct model is not estimable with the available data. Joiner [17] gave interesting examples.

At the end of the day, the risk of having missed the correct model must always be considered in any model selection process. Sensible reporting will highlight the limitations of the data. As a reader of a model selection report, one must always remember ‘caveat emptor.’

3.3. Step 1: Specify the maximum model

The first step for building a good enough model is to specify the maximum model, the model with the most covariates among all the models considered. Subsequently, any model considered can be created by deleting variables from the maximum model. The process we describe implicitly assumes that the maximum model contains a good enough model. More strongly, the process appears to imply that the maximum model contains the correct model. However, can we do good to science even when it does not? We believe that following and correctly interpreting the process described in the subsequent sections will typically lead to valid conclusions, even when the maximum model does not contain the correct model.

A correct interpretation requires clearly stating and discussing the implications of the limitations of the data relative to the original scientific goal. For example, a new instrument promises a quicker and more convenient way to measure forced vital capacity (FVC, maximum lung capacity), a measure of pulmonary function. A regression model predicting the standard measurement from the new one would need to include age, gender, and other health indicators, such as the presence of certain diseases. The limits of the sample characteristics impose limits on the generalizability of the data in many ways. Including only healthy males of ages 18–30 in the sample, limits the generality of the findings and the ability to specify a model general enough to fit the entire population. More subtly, a sample of FVC values from such a restricted population will typically be judged to follow a Gaussian distribution, whereas a sample from a sufficiently wide range of ages will be better modeled as log-normal (arguably the correct model). In such cases conscientious use of sensitivity analysis, especially assumption diagnostics such as outlier and heterogeneity evaluation, can identify many kinds of wrongness in a model. Assessing the value of adding polynomial terms provides an important tool for assessing the validity of the linearity assumption, and hence whether the maximum model contains the correct model. Joiner [17] described diagnostic strategies for discovering the presence of lurking variables, i.e. variables that matter that are not included in the analysis. Even if the assumption diagnostics do not highlight problems, poor fit (such as low R^2) will still emphasize the possibility of the correct model not being contained in the maximum model.

The selection of the maximum model depends on both statistical considerations and scientific considerations in the specific area. If from the scientific point of view some predictors play important roles relative to the outcomes, then the predictors should be included in the maximum model regardless of whether they are statistically significant or not in the model with the current data. From the statistical point of view, one could choose a large maximum model to avoid omitting a significant variable (avoid any Type II errors) and, therefore, maximize validity and predictive power. Alternately, one could choose a small maximum model to avoid including a non-significant variable (avoid any Type I errors) and, therefore, maximize reliability and parsimony while avoiding collinearity.

3.4. *Step 2*: *Specify a model selection criterion*

3.4.1. Single-model criteria. The second step for building a good enough model is to specify a model selection criterion. In practice, the coefficient of multiple determination R^2 is often used to measure the overall goodness of fit of an ordinary multiple linear regression model. The larger the R^2, the better the model fits the data.

For linear mixed models fitted with the two-stage approach, a subject-specific coefficient of multiple determination $R_i^2, i \in \{1, 2, \ldots, N\}$, can be used to assess the goodness of a first-stage linear model to the observed longitudinal profiles, where $R_i^2 = (\mathrm{SSTO}_i - \mathrm{SSE}_i)/\mathrm{SSTO}_i$. Verbeke and Molenberghs [10] suggested using scatter plots of the R_i^2 values versus the number of repeated measurements p_i to summarize the R_i^2, $i \in \{1, 2, \ldots, N\}$, and assess the goodness of fit of the model. Ideally, all R_i^2 should be large, say $\geqslant 0.75$. If the scatter plot of subject-specific values of R_i^2 versus the p_i shows that many R_i^2's are small, then one may consider a different first-stage model, e.g. change a linear to a quadratic model.

One might also use an overall measure to assess the goodness-of-fit of a first-stage linear model: $R^2_{\mathrm{meta}} = \sum_{i=1}^{N} (\mathrm{SSTO}_i - \mathrm{SSE}_i)/\sum_{i=1}^{N} \mathrm{SSTO}_i$ [10]. This measure indicates what proportion of the total within-subject variability can be explained by the first-stage linear model [10].

A better overall statistic was suggested by Edwards *et al.* [18]. They developed a model R^2 statistic for the linear mixed model based on an appropriate F statistic. The R^2 statistic measures multivariate association between the repeatedly measured outcome and the fixed effects and assesses the goodness of fit of the fixed effects model (the means model).

3.4.2. Multiple models comparison criteria. In practice, everyone prefers a simple model that fits the data well. Therefore, in addition to assessing the goodness of fit of a single model, we often compare different models to see if a simpler model fits the data nearly as well. Three types of mixed model comparisons can occur: (1) Compare mean models with the same covariance structure. Nested mean models are the most common. (2) Compare covariance models with the same mean structure. Two mixed models may be nested or non-nested in their covariance models. (3) Compare mixed models with different means and different covariance structures. When addressing fixed effects (mean model), we compare nested mean models with the same covariance structure. When addressing random effects (covariance structure), we recommend comparing covariance models with the same mean structure.

The most common comparison involves two nested models with the reduced model nested in the larger model. The two models differ only by the deletion of variables from the larger model, either explicitly or implicitly (through the imposition of constraints on the larger model). When parameters of the two models are estimated with the maximum likelihood (ML) method, a likelihood ratio (LR) test can be used to compare the reduced model with the larger model. With L_j the log likelihood for model j, the LR test statistic $T = -2(L_1 - L_2)$ asymptotically follows and therefore is referred to a $\chi^2_{\boldsymbol{d}}$ distribution, where d is the difference in the number of parameters between two models.

Alternatively, models can be fit by restricted maximum likelihood (REML), which is used in estimating covariance components by maximizing the likelihood function of a set of error contrasts (instead of maximizing the likelihood function of the data) to account for the loss of the degrees of freedom involved in the estimation of fixed effects. In contrast to ML estimation, when models are fit by REML, changing the fixed effects of the model lead to a different design matrix $\boldsymbol{X}$ and hence different error contrasts. Consequently, the corresponding REML likelihood functions are based on different observations and are not comparable. Therefore the difference in -2 restricted maximum log likelihood values between nested fixed effects models does not give a valid LR test [11]. Instead, alternative tests [19] should be used for testing nested mean structures (fixed effects) when REML is used for the model fitting.

In contrast to comparing nested mean models, when comparing nested models with different covariance structures but with the same mean structure (fixed effects), the REML likelihood functions from the two models are comparable because the same mean structure leads to the same error contrasts. Consequently a valid LR test can be obtained with the REML approach when comparing nested covariance structures where the true values of covariance parameters are not on the boundary of the parameter space.

Another situation in which a classical LR test cannot be used with either ML or REML estimation arises in some tests about covariance structure against zero. When testing the variance components against zero, the true values (**0**) of parameters under the null hypothesis are on the boundary of the parameter space. Consequently the LR test statistic under the null does not follow a chi-squared distribution asymptotically but often follows a mixture of chi-squared distributions [20, 21]. It, therefore, cannot be referred to a χ^2 distribution for the *p*-value. Instead, in the simple case of testing s versus $s+1$ random effects, the *p*-value can be computed as an average of *p* values with respect to χ^2_s and χ^2_{s+1}. Stram and Lee [22, 23] provided a detailed discussion on more complicated cases.

Various information criteria have been proposed as alternatives to LR tests in selecting models, especially when the models to be compared are not nested such that one cannot use an LR test to compare them as we discussed above. The main idea behind the information criteria is not to compare models with their maximized log likelihood values but to penalize for the use of too many parameters. The model with the smaller information criterion is usually preferred. However, note that the information criteria only provide some guidelines to compare models and should not be used as formal statistical tests of significance. Two commonly used information criteria are Akaike's [24] and Schwarz's [25] criteria. Akaike's information criterion (AIC) is more conservative and tends to choose more complex models than Schwarz's Bayesian information criterion (BIC) [26]. Therefore, AIC is preferable if Type I error control is the highest priority, but BIC is preferable if one would like to attain a higher power. The finite-population-corrected AIC (AICC, [27]) seeks a compromise between AIC and BIC in terms of loss of power. We agree with Morrell *et al.* [28] who stated 'The best way to select among linear mixed-effects models based on various information criteria is still not clearly determined.' In a similar vein, Wang and Schaalje [29] observed that in their evaluations 'Characteristics of the data, such as the covariance structure, parameter values, and sample size, greatly impacted the performance of various model selection criteria. No one criterion was consistently better than the others.'

3.5. *Step 3*: *Specify a predictor selection strategy*

The third step for building a good enough model is to specify a predictor selection strategy. Predictor selection strategies include:

- All possible regressions strategy, which requires fitting all possible models and therefore provides the most thorough approach for exploratory model selection. However, the computational burden is its primary disadvantage and challenge.
- Backward elimination, which begins with the maximum model and delete variables of no value.
- Forward selection, which begins with the simplest model and adds variables of most value.
- Stepwise, which begins with a forward selection step and allows a deletion step after each addition.

We note that in mixed models, unlike in ordinary linear models, the outcome of the predictor selection strategy also depends on the covariance structure used. Therefore, finding a good enough covariance model is important for the model selection of the fixed effects. Section 4 provides practical advice on the selection of predictors and covariance structures. As for expected value (fixed effect) estimation, overfitting yields unbiased estimates, although it loses a few degrees of freedom. However, underfitting can create severe bias, which does not diminish with the increased sample size. Therefore, among strategies, we recommend beginning with the maximum model and then conducting backward elimination to delete variables of no value. Nilsson *et al.* [30] showed that backward elimination is consistent given any strictly positive (non-zero) distribution for covariates. The assumption of strictly positive distribution for covariates is considered reasonable whenever there is uncertainty about the data, i.e. data are obtained with noise [31]. It should be emphasized that the statistical tests commonly used in mixed models (*F*, *t*-test) are all approximate. Hence, caution should be the rule when assessing effects that may be marginal. In addition to statistical considerations, scientific considerations should also play an important role in the predictor selection. If some predictors are important from a scientific point of view, then they should be kept in the model even though they are not statistically significant with the current data.

3.6. *Step 4*: *Conduct the analysis*

After we define the maximum model and have chosen our criteria for model reduction, we are ready to conduct the analysis. In the process of the analysis, collinearity should be assessed and eliminated from models. In addition, assumption diagnostics and influence diagnostics should be performed to check the validity of the model assumptions and the impact of individual observations on the analysis. These diagnostics are often helpful in the model selection. In Section 4, we will provide more detailed and practical advice on collinearity and diagnostics. Muller and Fetterman [7] provided advice on issues for ordinary linear models. Their strategies are also helpful for mixed effects models.

3.7. *Step 5*: *Evaluate the reliability*

Once the final model is selected, we would like to evaluate whether the model holds when being applied to a new sample (since we are often interested in predicting the outcome for a new sample). We suggest using a split-sample analysis to evaluate the reliability. That is, before the analysis begins, we randomly split subjects to either the training data for exploratory analyses or the holdout data for evaluation of reliability, where all observations for a given subject will be in the training data or the holdout data.

All exploratory analysis is restricted to the training sample, while the holdout sample is reserved and not used in any way during the exploratory phase. The holdout data are retained for subsequent use in confirming and validating the initial analysis. The process must be distinguished from what some statisticians and scientists describe as a 'cross-validation' in which the data are recombined and the splitting and analysis process is repeated. The splitting process must be statistically independent of the response or outcome values but could be stratified with covariate values.

After the splitting, we perform the following procedures. Note that each subject has several outcome observations over time in a longitudinal study. Therefore, in the following procedure we recommend using the overall R^2 proposed by Edwards *et al.* [18] as a measure of the squared correlation between the observed and predicted outcomes in the linear mixed models. For generalized mixed models, we recommend using the Euclidean distance between the observed and predicted outcomes to replace the squared correlation discussed below.

- Conduct exploratory analyses on the training data based on Steps 1–4 and select an exploratory model.
- Apply the exploratory model obtained from the training data to the holdout data, and then compute corresponding statistics with the holdout data and compare them with statistics computed with the training data.
  - For the training data, compute the squared correlation between the observed outcomes and the predicted outcomes based on the model selected in Step I from the training data.
  - For the holdout data, compute the corresponding statistic (called the squared cross-validation correlation) between the observed outcomes and the predicted outcomes based on the model selected in Step I from the training data.
  - Compare the squared correlation from the training data and the squared cross-validation correlation from the holdout data. The difference between these two correlations is called shrinkage on cross-validation. The closer the two correlations are, and therefore the smaller the shrinkage the better, indicating a higher reliability of the model.
- Combine training and holdout data and conduct regression diagnostics on the pooled data.
  - If shrinkage is small (i.e. reliability is high), pool the training and holdout data to provide the best available estimates of parameters and diagnostics.
  - If shrinkage is poor, pool the data to conduct a second-round exploratory analysis and overall diagnostics.
- Report the results. This includes all results in steps I, II, and III, including any failure to cross-validate and subsequent reanalysis.

In this section, we gave a framework and strategy on fitting a good general model. In the following section, we will give specific and practical advice on fitting a good enough mixed model, including how to construct mean and covariance structures, how to diagnose and treat for collinearity and model assumptions, and how to help model convergence.

4. Practical advice

4.1. Diagnostics and treatments for collinearity, convergence, computing speed, and numerical accuracy

For an otherwise valid model, many important statistical properties are not affected by collinearity. Examples include error variance and correlation estimates in an otherwise valid univariate model, and the distributions of residuals. However, collinearity does affect estimation and inference for the primary parameters, causes a loss of power, and makes interpretation far more difficult. When there is collinearity, the estimates of regression coefficients are unstable and imprecise, iterative programs are less likely to converge or will take longer to converge, results can depend on the listing order of variables, and adding or dropping a few cases may greatly change the estimates of coefficients. Therefore, we recommend avoiding collinearity in fitting models by scaling (e.g. use 1–2 m for height rather than 100–200 cm) and mean centering (i.e. use $X-\bar{X}$ rather than X) or roughly centering predictors (not exact mean centering but subtracting some 'nice' number to simplify interpretation). Centering and scaling can substantially reduce collinearity. It would be hard to overemphasize the value, in practice, of the simple strategies of centering and scaling as a first step in improving convergence, computing speed, and numerical accuracy.

With well-centered and scaled data, replacing less-than-full-rank coding for predictor variables with full rank coding can greatly improve the convergence, computing speed, and numerical accuracy. Furthermore, using full rank coding also helps in reducing collinearity and hence improve the interpretability and statistical reliability. In full rank coding schemes, the number of columns of the design matrix in the model equals to its rank, and the columns of the design matrix are linearly independent of one another. The full rank coding schemes include cell-mean coding (i.e. create G indicator variables for G groups and do not create an intercept), reference-cell coding (i.e. create an 'intercept' and $G-1$ indicator variables for G groups, where indicator variables take values 0 or 1) or effect coding (i.e. create an 'intercept' and $G-1$ variables for G groups, but those $G-1$ variables take values of -1, 0, and 1). For example, suppose we have four groups A, B, C, and D. Then, in the cell-mean coding scheme, we will not create an intercept but have four indicators A, B, C, and D to indicate the group membership, where one observation takes the value 1 if it belongs to the group and 0 otherwise. In the reference-cell coding scheme, we will have one intercept (for the reference group A) and three indicators for B, C, and D, where observations in group A will take the value 0 for all the indicators B, C, and D. Effect coding differs from reference-cell coding only by setting the 0 values to -1 only for all observations in group A, for indicator variables B, C, and D. Muller and Fetterman [7] gave more detailed information on different coding schemes.

Compared with the ordinary linear regression, collinearity in mixed models may stem from either the fixed or random effects. Stinnett [32] had a detailed discussion on diagnostics and treatments of collinearity in mixed models to avoid inflated variances of covariates' coefficients in fixed effects. For that purpose, Stinnett [32] suggested assessing collinearity after the estimation of model parameters because the variances of the coefficients involve the covariance structure.

In this paper, we provide advice on collinearity that can be easily implemented and addresses the problem reasonably in practice. Basically collinearity is a problem among predictors; hence, it is reasonable in practice (although not necessarily perfect) to assess collinearity among predictors in ordinary regression models and software separately from fitting mixed models. As discussed by Muller and Fetterman [7] in Chapter 8, all widely used measures of collinearity are functions of the design matrix, $\boldsymbol{X}$, through the cross-products matrix $\text{SSCP}=\boldsymbol{X}'\boldsymbol{X}$ or the corresponding covariance or correlation matrix (among predictors). A condition index (CI) can be computed from ratios of eigenvalues for any of the three cross-product matrices (SSCP, covariance, correlation), whereas the corresponding eigenvectors provide variance decomposition information to identify problem predictors. When it exists, the inverse correlation matrix has diagonal elements known as variance inflation factors, (VIF_j), with $\text{VIF}_j=1/(1-R_j^2)$ and R_j^2 is the squared correlation of predictor j with the remaining $q-1$ predictors ($q-2$ if the design matrix $\boldsymbol{X}$ includes an intercept). The value of $(1-R_j^2)$ is often called the tolerance. CI of 5 to 10, 30 to 100, and >100 suggest weak, moderate to strong, and serious dependencies, respectively [33] (Appendix A1 provides the SAS code.). In practice, the tolerance or variance inflation factor (VIF) is useful in detecting overall collinearity problems and can be easily obtained from the most standard software (Appendix A1 provides the SAS code.). Usually a small tolerance (<0.1) or high VIF (>10) indicates a collinearity. However, the validity of the evaluation depends on the ability to accurately compute the VIF values, which becomes impossible if collinearity increases sufficiently. Hence we recommend first examining the less commonly used CI because they always exist and can usually be computed accurately even with extreme collinearity.

When collinearity is diagnosed, the common treatments for it include eliminating predictors with near-zero variance, eliminating redundant variables, redefining or combining variables, and using cell-mean coding, reference-cell coding, or effect coding. Effect coding provides the best choice for controlling collinearity among the coding schemes [7].

4.2. Selection of a preliminary mean structure, random-effects structure, and correlation structure

4.2.1. Overall approach. As discussed above, although the mean structure (the fixed effects) is usually of primary interest, an appropriate covariance model is important to obtain valid inferences for the fixed effect parameters. Therefore, unlike ordinary regression models, fitting mixed models implies that both an appropriate mean structure and covariance structure need to be specified. As we discussed in the five steps in Section 3, like ordinary regression models, we recommend starting with a maximum model in both fixed and random components and deleting redundant terms to avoid underfitting bias. In addition, guidelines for fitting classical regression models should be applied to mixed models too. For example, all hierarchically inferior terms should be included if a higher-order effect is included.

4.2.2. Selection of a preliminary mean structure. The selection of covariates in the fixed effects is the same process as the selection of predictors in the ordinary regression models, and therefore is based on both scientific and statistical considerations. Variables considered to be important from the scientific point of view should be included as covariates in the mean structure. In contrast to ordinary regression models, for the longitudinal data we also need to decide what functions of time should be included in the mean structure model. The science of the situation will imply the order of the model likely needed. For example, a plausible model of child growth may require up to a cubic polynomial or restricted cubic splines to account for an initial level, a growth phase, and a plateau. It is most important at the preliminary stage to avoid specifying a model that is too small and therefore does not contain the correct model. The later stages of model fitting allow for reducing the model to the simplest that suffices. Either planned step down tests or thorough exploratory analysis, such as plots of smoothed average trend or individual profiles over time, may be used, depending on the strategy selected (i.e. confirmatory or exploratory). For example, if the smoothed average trend or individual profiles suggest modeling the outcome as a quadratic function over time, then an intercept and both a linear and a quadratic time effect (time, time^2) should be included in the mean structure. If we do not want to make any assumptions about the shape of the curve, time may be included as a categorical predictor in the fixed random effects.

4.2.3. Selection of a random-effects structure and correlation structure. In parallel to defining the fixed effects model, a random effects model must be chosen to define a covariance model. After we decide the fixed effects, we need to select a set of random effects to be included in the model. We agree with Verbeke and Molenberghs [10] who wrote 'Note that random effects for time-independent covariates can be interpreted as subject-specific corrections to the overall mean structure. This makes them hard to distinguish from the random intercepts. Therefore, one often includes random intercepts, and random effects only for time-varying covariates.' In the selection of random effects for time-varying covariates, we usually only consider time-varying covariates that have been included in the fixed effects (mean structure).

During the actual data analysis, as discussed in the subsequent sections, a plot of the ordinary least-squares (OLS) residual profiles versus time is helpful for the selection of random effects during exploratory model fitting. For example, if the plot shows constant variability over time, then only random intercepts are included; whereas if the plot suggests some straight lines, then random intercepts and random slopes for time are both included (see Section 5 for illustration with an example). For other random effects, Verbeke and Molenberghs [10] suggested an informal check for the appropriateness of the selected random effects in Section 9.3.

Next, we need to select an appropriate covariance or correlation structure for the model to account for the correlation among measurements. Regarding the response variable's marginal covariance structure, the overall structure $\boldsymbol{\Sigma}_i$ is the sum of the random effects portion $\mathbf{Z}_i\Sigma_{\mathbf{d}_i}(\tau_{\mathbf{d}})\mathbf{Z}_i'$ and the residual error portion $\Sigma_{\mathbf{e}_i}(\tau_{\mathbf{e}})$. In cases when the variability in the measurements cannot be completely modeled by the random effects, we use both random effects and residual errors to describe the covariance structure. However, in practice, it is often efficient to use one of the two common ways to model the covariance or correlation structure: (1) Model the covariance structure only through the random effects (use the *Random* statement in SAS Proc Mixed or Proc Glimmix) and assume conditional independence of the residuals, i.e. given the random effects, the covariance for the residuals is given by an unknown scalar (variance) multiplied by an identity matrix; (2) Do not include random effects but model the covariance structure only through the residual errors (use the *Repeated* statement in SAS Proc Mixed or Proc Glimmix). However, the latter approach technically would no longer qualify as a mixed effects model. Commonly used covariance structures through the random effects include compound symmetry (CS) for constant correlation between any pair of repeated measurements, first-order autoregressive (AR(1)) for stronger correlation among adjacent measurements, and general unstructured (UN) covariance model for non-specific correlation among measurements. Verbeke and Molenberghs [10] discussed specific residual covariance structures.

Note that when we use both random effects and residual errors (use both *Random* and *Repeated* statements in SAS Proc Mixed or Proc Glimmix) to describe the covariance structure, the components typically do not add up to some known covariance structure. For example, an UN random effects covariance (say for a random intercept and slope) and an AR(1) residual error covariance lead to a covariance of the response that is not structured in the way that we speak of above. With repeated measures, it is often appropriate to assume a relatively simple random-effects structure for covariance and a more complex structure for the residual error covariance. If we do not want to make any assumptions about the shape of the curve, time may be included as a categorical predictor in the fixed and random effects, which corresponds exactly to a fully saturated polynomial design matrix in time. In such cases, a wide range of covariance structures has appeal, especially UN patterns.

Similar to the selection of fixed effects, the evaluation of covariance structure should depend on scientific knowledge in the specific area as well as statistical criteria. Covariance structures inconsistent with the context of the data should be ruled out. However, due to the need to use iterative procedures for computing estimates, convergence can become an issue. In most cases, when a mixed model algorithm does not converge, it is typically assumed to be due to difficulties with the estimation of the covariance parameters. However, our experience leads us to believe that many, and perhaps most, such difficulties disappear with careful use of (1) centering, (2) scaling, (3) full-rank coding, and (4) collinearity removal for predictors in both X and Z.

In the context of analysis planning, the logic of the sampling plan typically suggests a particular class of covariance structures. In contrast, in the context of data analysis, as distinct from analysis planning, the data in hand may not support the logically obvious structure. It would be ideal to rule out covariance structures inconsistent with the data. In practice most analysts chose to simplify the covariance structure to achieve convergence, which leads to the risk of underfitting and the consequent risk of biased inference. The estimated covariance and correlation matrices and plots of individual profiles can help to identify any systematic covariance structure. Without obvious systematic structure, one usual approach is to start with a general UN covariance structure, and then fit the same model again with a simpler covariance structure, e.g. CS, and compare the two models with difference in likelihoods and information criteria.

The current statistical practice appears to center on avoiding the problem by simplifying the covariance model. We urge the reader faced with the problem to consider a three-step response. First, step back to recheck scaling, centering, and collinearity, and examine the patterns of missing and mistimed data. The collinearity check must be based on full-rank coding with the dummy variables for the categorical predictors included in the design matrix. Second, if necessary, modify the data (which may lead to simplifying the model) as needed to improve convergence and try again. Third, when compelled to fit a reduced covariance model (inconsistent with the planned structure) include explicit discussion on the risks to accurate inference (due to inaccurate estimates of variances) in any report of the analysis.

The choice of covariance model can lead to difficulties in a variety of ways. Although we favor including more rather than fewer fixed and random effects, it should be noted that overfitted models may result in divergence of the maximization procedure. Some mixed model analyses, especially in small samples, can give inflated type I error rates and confidence intervals that are too narrow. The problem could be due to any combination of an insufficiently complex and untested (and therefore not credible) covariance structure, insufficient data for the scientific goal, an inadequate model selected for fitting, inadequate knowledge about covariance structures when using mixed model software, and computationally fragile software.

4.3. Assumption diagnostics

In the selection of mixed models, the validity of the model assumptions should be checked with the maximum model and reduced models. The nature of mixed models requires some special thinking about assumptions, but need not require much more work. We suspect many neglect checking assumptions due to assuming that distinct and special methods must be used. Below we sketch examples that demonstrate that for the linear mixed models with the Gaussian errors, the better-known tools from the univariate models can be applied to great effect.

For the linear mixed effects models, assumptions of normality, linearity, and variance homogeneity should be assessed with residuals. Note that one persistent error in data analysis is using raw data rather than residuals to assess distribution assumptions. There are two kinds of residuals in a mixed model. One is a marginal residual, which is a deviation of a person from the group mean; another is a conditional residual, which is a deviation of one measurement within person from the mean of that person over time. Although both residuals are useful, one needs only to test the normality with marginal residuals [34].

A jacknifed studentized residual involves the difference between a subject's observed outcome value and his/her predicted value based on the data set in which that case is deleted. It has some nice properties, e.g. it exactly follows a t distribution and its distribution corresponds to the distribution of a predicted future observation (see [24] for detailed discussions). Therefore, we recommend using jackknifed studentized residual histograms and box plots, and scatter plots of jackknifed studentized residuals versus predicted values over time to help the assessment. If the plots suggest some problem of non-normality and heterogeneity, then one should consider transformations like the Box–Cox transformation as generalized to the mixed model by Gurka *et al.* [34], or use generalized mixed models or heterogeneity models if necessary. If linearity is a problem, then try to find useful predictors and transformations, or use nonlinear models if necessary.

4.4. Improve convergence

When fitting mixed models, one may have problems getting the model to converge. There are many reasons for a model failing to converge, including too small a sample, the presence of collinearity (which may reflect any combination of poor centering, scaling, or dummy coding), a correctable mistake in the choice of error distribution, and a complicated covariance structure like high-dimensional random effects with an unconstrained covariance structure. However, as stated previously, in most cases with good centering, scaling, coding, and limited collinearity, when a mixed model algorithm does not converge, it is due to estimation of the covariance parameters.

When the maximization procedure diverges, one can first try some computational treatments such as increasing the maximum iterations, changing the parameter convergence values for step-halvings, the singularity tolerance value, or the number of scoring steps. If they do not help, one can try simplifying the data by rounding time to the largest units that have reasonable scientific meaning (e.g. use monthly data instead of weekly data) or use neighborhood averages. Without violating the credibility, one can also simplify the mean structure and covariance structure of the model by removing redundant variables, avoiding high-order terms, or selecting a simpler covariance structure (while recognizing the risks of underfitting).

4.5. Recommendations

Finally, we would like to give some overall recommendations, as follows, in selecting a good enough mixed model:

- Plan, assess, and report analysis openly.
- Plan backwards by considering the following steps:
  - What are the research goals?
  - What is the outcome distribution?
  - Start with the maximum model by selecting an adequate preliminary mean structure and random-effects structure.
  - Implement model reduction.
  - Select the final model and do final analysis.
  - Conduct diagnostics.
  - Report results.
- Avoid, assess, and eliminate collinearity.
- Check assumptions carefully.
- Specify and evaluate a credible covariance model.
- Use transformations and non-Gaussian forms, e.g. use models for a non-Gaussian exponential family.
- Fit models backwards.
- Provide all details of what you actually did in the written report.

5. A real example

5.1. Overview of the real example

In this section, we will use a real example to illustrate the strategies discussed in this paper. The data are from a group-randomized controlled trial ($N=5812$) of a multicomponent alcohol preventive intervention for multiethnic urban youth, called Project Northland Chicago [35–37]. Sixty-one public schools in Chicago that included grades 6–8 were matched on ethnicity, poverty, mobility, reading, and math test scores and randomly assigned to either the intervention or control group. The intervention was to have teachers, peer-leaders, parents, and community members collaborate with students to prevent alcohol drinking. Students were followed at 0.5, 1.5, and 2.5 years after enrollment. For illustrative purposes, in this paper we focus on the association of covariates with alcohol use in the last 30 days assessed with item ‘During the last 30 days, on how many occasions, or times, have you had alcoholic beverages to drink?’ The subsample for the present study comprised 1388 students from 34 schools and their parents who were present at baseline and were part of the control group. A substantive application of this example can be found in Komro *et al.* [37]. Tables I and II show the baseline characteristics and outcome distributions at baseline and three follow-up times.

5.2. Model the outcome as a continuous variable?

The outcome of interest is alcohol use in the last 30 days at 0.5, 1.5, and 2.5 years after enrollment, for which one common way to model is to treat it as a continuous outcome (number of occasions in the last 30 days). In this subsection, we will treat

Table I. Descriptive statistics for covariates.

Variable	Mean	SD	Percentage of 'Yes'
Home access at baseline			
Last time drank, received from parent	—	—	9.87
Last time drank, took from home	—	—	1.01
How hard to get alcohol from parent[*]	2.83	0.46	—
How hard to get alcohol from home[*]	2.59	0.69	—
Done things to make home access more difficult	—	—	45.19
Parent report sixth grader allowed to drink alcohol in home	—	—	5.91
Ask sixth grader to bring alcohol to you, past 30 days	—	—	5.49
Home alcohol access scale[†]	8.33	4.35	—
Covariates at baseline			
Student			
African American	—	—	44.46
Hispanic	—	—	39.19
White	—	—	16.35
Age at baseline	11.83	0.57	—
Gender (male)	—	—	48.38
Language spoken most often at home			
English	—	—	67.23
Time lived in US[‡]	1.26	0.74	—
Receive reduced-price lunch	—	—	72.43
Family composition (mother and father, together)	—	—	55.91
Parent/child communication scale[§]	21.04	4.42	—
Family alcohol discussions scale[¶]	10.56	3.51	—
Number of friends who drink[‖]	1.49	0.85	—
How hard to get alcohol from friend[*]	2.34	0.78	—
Parent			
Parental monitoring scale[**]	32.86	4.19	—
Alcohol communication scale[††]	19.12	5.66	—

[*] 1 = easy, 2 = in between, 3 = hard.
[†] Range: 4–20; a higher score on this scale indicates more access.
[‡] 1 = all your life, 2 = 7–9 years, 3 = 4–6 years, 4 = 1–3 years, 5 = *less* than one year.
[§] Range: 6–30; a higher score on this scale indicates more parent/child communication.
[¶] Range: 3–15; a higher score on this scale indicates more family alcohol discussions.
[‖] 1 = none, 2 = a few, 3 = some, 4 = many, 5 = almost all.
[**] Range: 11–40; a higher score on this scale indicates more parental monitoring.
[††] Range: 5–25; a higher score on this scale indicates more communication about alcohol.

the outcome as a continuous variable and fit a linear mixed effects model to the data, then conduct diagnostics to examine the validity of the model assumptions. Sections 5.2 and 5.3 will show model fitting on transformed outcomes.

Based on the scientific considerations, the outcome at baseline and 21 variables were included as covariates in the maximum model, including characteristics of subjects and alcohol access variables (Table I and [37] have detailed information).

After checking and adjusting scaling of the data, to avoid collinearity, we fitted a linear regression model on the outcome with the 21 covariates and then conducted collinearity diagnostics on them (Appendix A1 provides the SAS code). Since the tolerances were relatively small for the model, we deleted four variables based on scientific understandings of the problem. This study was conducted in homogeneous low SES communities. The variables 'Language spoken most often at home', 'Time lived in US', and 'Receive reduced-price lunch' were highly correlated with race, and the variable 'Done things to make home access more difficult' was highly correlated with another variable 'How hard to get alcohol from home', so we only kept the variables 'Race' and 'How hard to get alcohol from home' but deleted four other variables to avoid collinearity. Furthermore, we recentered variables 'age at baseline' and 'parental monitoring scale' by subtracting their approximate means 12 and 32, respectively; therefore, we included in the model 'age at baseline—12' and 'parental monitoring—32'. As discussed previously, we suggest including more fixed effects rather than fewer to build a model based on both statistical and scientific considerations. Therefore, all other variables, which were all scientifically important to and measured different components of factors in alcohol use were included as covariates in the mean structure model. To explore how to model the outcome over time as a fixed effect, we plotted the average number of alcohol uses in the last 30 days versus time (Appendix A2 provides the SAS code). Figure 1 shows that after baseline, the outcome increases approximately linearly over time, indicating that modeling the outcome as a linear function over time in addition to including the baseline as a covariate in the mean structure model is appropriate for the data.

Table II. Outcome distributions at baseline and follow-up time points.

	Time			
Alcohol use in the last 30 days*	Baseline frequency (per cent)	0.5 years frequency (per cent)	1.5 years frequency (per cent)	2.5 years frequency (per cent)
1	1268 (91.75)	1137 (88.83)	783 (82.25)	664 (74.69)
2	93 (6.73)	110 (8.59)	122 (12.82)	156 (17.55)
3	16 (1.16)	22 (1.72)	24 (2.52)	44 (4.95)
4	3 (0.22)	4 (0.31)	11 (1.16)	9 (1.01)
5	1 (0.07)	4 (0.31)	5 (0.53)	11 (1.24)
6	0	1 (0.08)	2 (0.21)	1 (0.11)
7	1 (0.07)	2 (0.16)	5 (0.53)	4 (0.45)
Total	1382	1280	952	889

*1: no drink; 2: 1–2 occasions; 3: 3–5 occasions; 4: 6–9 occasions; 5: 10–19 occasions; 6: 20–39 occasions; 7: ⩾40 occasions.

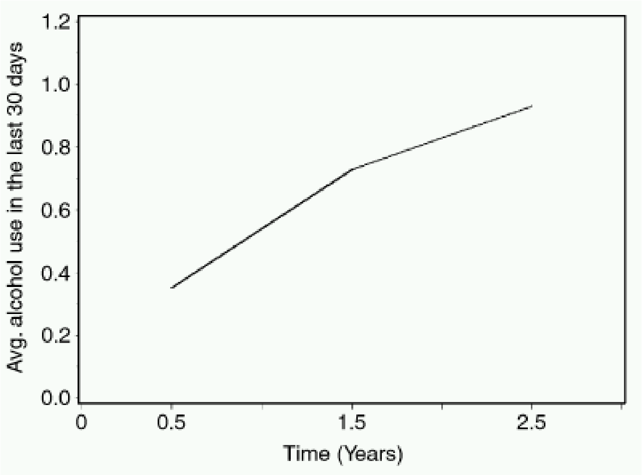

Figure 1. The average number of alcohol use in the last 30 days over time.

After the exploratory selection of fixed effects, to select random effects we restricted our attention to variables already included in the fixed effects/mean structure model. In this study the only time-varying variable is time itself. To choose appropriate random effects for time, we plotted residual profiles versus time in Figure 2 (Appendix A3 has the SAS code), which used jackknifed studentized residuals discussed in Section 4.3. From Figure 2, we see constant variability over time, indicating that no random effect is needed to account for variability over time and therefore it is appropriate to include random intercepts only. In addition, since students were nested within schools, we specified schools as a nested random effect to account for the design effect. Including random effects for other time-independent variables is not of interest in this paper; hence, no other random effects were included. To select a covariance structure, we started with a general UN covariance matrix to avoid problems due to an inappropriate covariance structure. The linear mixed effects model was fitted with the mean structure and covariance structure discussed using SAS (Appendix A4 provides the SAS code.).

We next needed to check the validity of the model assumptions. We used the marginal residuals to assess the normality assumption with the data. The jackknifed studentized residual histograms over time are skewed (Figure 3 shows the histogram at 1.5 years for illustration.) and the Kolmogorov–Smirnov normality test on the residuals gave a value of $D=0.315(p<0.01)$, indicating that there is a strong evidence that the normality assumption does not hold for the data.

5.3. *Model the outcome transformed?*

Because of invalidity of the normality assumption, we tried several transformations of the outcome, including square root, log, and Box–Cox transformations with power from 2 to 16. With each transformed outcome, we refitted a linear mixed effects model with the same mean and covariance structures as the model in Section 5.1. However, with all the transformed outcomes, the jackknifed studentized residual histogram and normality test still indicated that the normality assumption does not hold.

Alternatively, one can try a two-part (mixture) model, in which one component models the probability of any alcohol use and another component models the amount of alcohol use given that a student uses some alcohol. In this example, very few students had alcohol use for more than 3–5 occasions in the last 30 days (see Table II), which requires a more complex model in the longitudinal setting for the second component. Furthermore, interestingly, two residual plots (Figures 4 and 5) from the linear

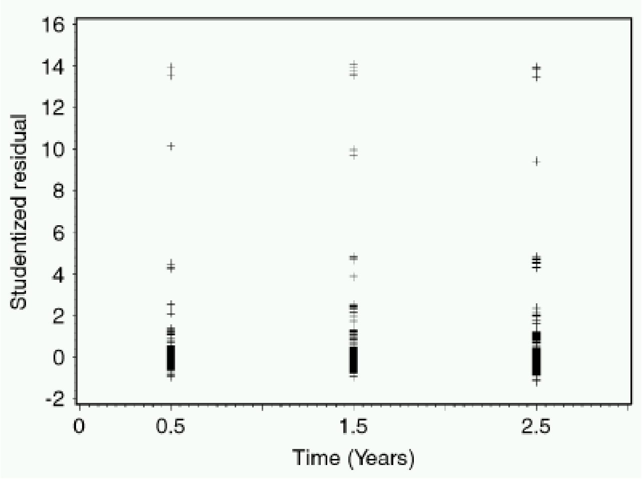

Figure 2. The OLS residual versus time.

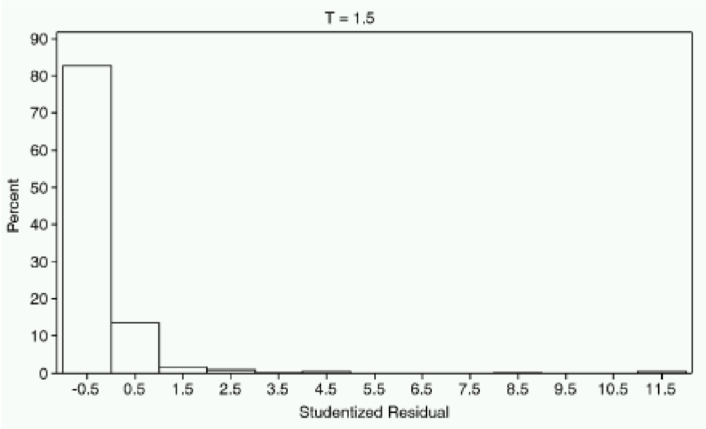

Figure 3. The Jackknifed studentized residual histogram from the linear mixed effect model with the original outcome at 1.5 years.

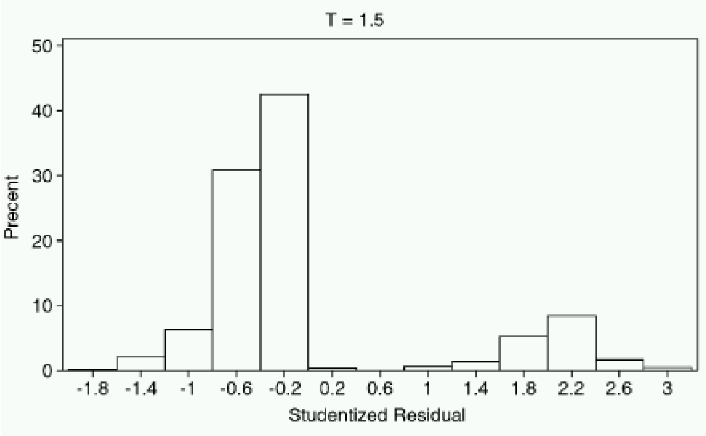

Figure 4. The Jackknifed studentized residual histogram from the linear mixed effect model with the Box–Cox (power $=16$) transformed outcome at 1.5 years.

mixed effects model with the Box–Cox (power $=16$) transformed outcome show an obvious pattern of two modes, indicating that recoding the outcome as a binary variable could be appropriate. The scientific focus of the work led the collaborators to choose the simpler route of focusing on the outcome as a binary variable.

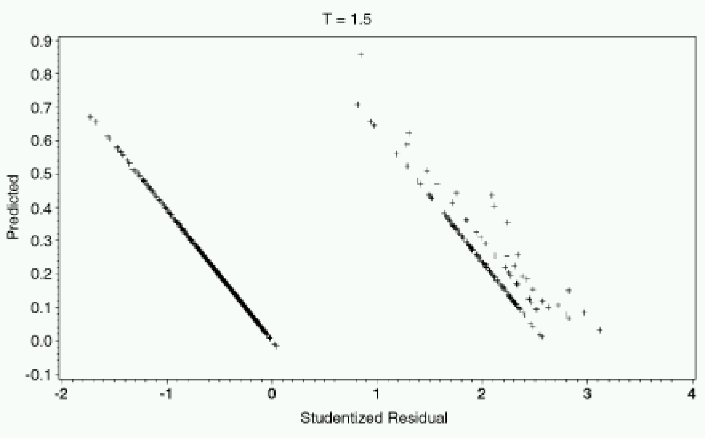

Figure 5. Predicted value versus Jackknifed studentized residual from the linear mixed effect model with the Box–Cox (power =16) transformed outcome at 1.5 years.

Table III. Association between variables at age 12 and trajectories of alcohol use in past month from age 12 to 14.

Variable	Slope	SE	OR	95 per cent CI
Intercept	−2.197‡	0.596		
Time	0.590‡	0.058	1.80	(1.61, 2.02)
Age at baseline (original scale −12)	0.314†	0.116	1.37	(1.09, 1.72)
Baseline	0.609†	0.204	1.84	(1.23,2.74)
Home access				
From parent survey				
Sixth grader allowed to drink at home	0.286	0.238	1.33	(0.84, 2.12)
Ask sixth grader to bring alcoholic beverage to you	0.143	0.251	1.15	(0.71, 1.89)
Home access scale	0.015	0.016	1.02	(0.98, 1.05)
From student survey				
Last time drank, received from parent	0.843‡	0.176	2.32	(1.64, 3.28)
Last time drank, took from home	0.583	0.562	1.79	(0.59, 5.39)
How hard to get alcohol from parent	0.020	0.135	1.02	(0.78, 1.33)
How hard to get alcohol from home	−0.232*	0.092	0.79	(0.66, 0.95)
Covariates				
Demographics				
Hispanic (ref: black)	0.565†	0.199	1.76	(1.18, 2.62)
White	0.450	0.245	1.57	(0.97, 2.55)
Gender (ref: female)	−0.272*	0.127	0.76	(0.59, 0.98)
Family composition	0.060	0.137	1.06	(0.81, 1.39)
Other				
Number of friends who drink	0.480‡	0.072	1.62	(1.40, 1.86)
How hard to get alcohol from friend	−0.159	0.086	0.85	(0.72, 1.01)
Parent/child communication	−0.033	0.018	0.97	(0.93, 1.00)
Parental monitoring scale (original scale −32)	−0.012	0.016	0.99	(0.96, 1.02)
Family alcohol discussions	−0.008	0.024	0.99	(0.95, 1.04)
Alcohol communication	0.005	0.013	1.01	(0.98, 1.03)

*$p<0.05$.
†$p<0.01$.
‡$p<0.001$.

5.4. *Model the outcome as a binary variable*

Since the normality assumption failed with both original and transformed outcomes, and Figures 4 and 5 show a pattern of two modes, we decided to transform the outcome to a binary variable. This transformation is scientifically meaningful for this study. We let the outcome be ‘0’ indicating no alcohol use in the last 30 days and ‘1’ otherwise.

One choice for a longitudinal binary outcome models population averages and finds estimates by solving generalized estimating equations (GEE) [38]. The second approach uses GLMM that includes a subject-specific component as well as the mean value of individuals. When one is interested in what happens 'on average' in the population, the population-averaged model is preferred; when the 'typical' (mean) value of individuals in the population and/or between-subject heterogeneity are of interest, then the subject-specific model is desired. Given our focus on how to build a good enough mixed model, we would not have a comparison between the applications of the two approaches in this example.

We fitted a generalized linear mixed effect model for the recoded binary outcome with the same mean and covariance structures used in Section 5.1 (Appendix A5 contains the SAS code for Proc Glimmix.). The process of selecting random effects was the same as in the linear mixed model. We started with an UN covariance structure and then compared the fit of several other covariance structures via comparison of likelihoods and information criteria. Finally an UN covariance structure was selected because of a substantial reduction in the likelihood when compared with any other covariance structures. Since the purpose of this example was not to find the most reduced model, we ended with this final model, which converged quickly. To find a reduced model, one can delete variables with the biggest *p* values and variables that are not important from the scientific point of view and proceed to fit the generalized linear models by testing the difference in log likelihoods. We also wished to avoid the burden of demonstrating replicable results needed for any extensive model reduction process.

The estimates and standard errors of coefficients and the odds ratio (*OR*) and 95 per cent confidence interval (*CI*) for each covariate were computed with the fitted model (see Table III). The results show that the past month alcohol use significantly increased over time ($P<0.001, OR=1.80(1.61, 2.02)$) and age ($P<0.01, OR=1.37(1.09, 1.72)$). Overall, 19 per cent of students reported having access to alcohol at home and students who reported any access to alcohol from home or from their parents had a significantly increased trajectory of the past month alcohol use, $OR=2.3(1.75, 2.97)$. A detailed discussion of these results can be found in Komro *et al.* [37].

6. Conclusions

In practice, investigators often ask how to build a good enough mixed effects model. We described a systematic step-by-step strategy including practical advice for achieving success. The advice we provide is convenient and easy to implement in practice with the available mixed model procedures in software such as SAS, SPSS, SPLUS/R, STATA, and others.

Appendix A: SAS code used in fitting models in the real example

```
*A1: Collinearity diagnostics with linear regression
proc reg data=ONE;
model Y=Ybaseline Time X1 X2.../collin tol vif;* Output collinearity, tolerance and VIF
run;quit;
*A2: Plot the average outcome over time for the selection of fixed effects for time
proc means data=ONE mean;
var Y; by time;
output out=Yout mean=avg_Y;
run;

symbol i=join color=black r=38;
proc gplot data=Yout;
plot avg_Y*time/nolegend vminor=1hminor=0;
run;quit;
*A3: Plot the OSL residuals over time for the selection of random effects for time
proc reg data=ONE;
model Y=Ybaseline Time X1 X2...
output out=resdout rstudent=r_i_Y;* Output Jacknifed studentized residuals
run;

proc gplot data=resdout;
plotr_i_Y* time;
run;quit;
*A4: Fit a linear mixed effect model and check residual plots
proc mixed data=ONE noclprint maxiter=200covtest;
class CLUSTER TIM ID;
model Y=Ybaseline Time X1 X2.../soutp=RES_Yresidual;
random INT/SUBJECT=CLUSTER;
repeated tim/type=UN SUBJECT=ID(CLUSTER);
run;quit;
```

```
*To check residual plots
proc univariate data=RES_Y normaltest;
var StudentResid;
by Time;
histogram;run;

proc gplot data=RES_Y;
plot pred*StudentResid;
by Time;
run; quit;
```

A5: Fit a generalized linear mixed effect model

```
proc glimmix data=ONE;
class CLUSTER TIM ID;
modelY=Ybaseline Time X1 X2.../s dist=binomial or outp=RES_Y residual;
random INT/SUBJECT=CLUSTER;
random tim/type=UN SUBJECT=ID(CLUSTER);
run;quit;
```

Acknowledgements

The paper is based on a presentation with the same name at the June 2007 meeting of the National Institute of Aging (NIA) Longitudinal Data on Aging (LDA) Working Group. We thank the National Institute of Aging for sponsoring the workshop and our participation. We also wish to thank David Allison and the other members of the organizing committee for inviting us to such a stimulating event. We would also like to thank the editor and the two referees for helpful comments that improved this paper. Cheng's work was supported in part by NIH/NIA grants P30-AG028740 and R01-AG026556-01A2. Muller's work was supported in part by NIH/NIDCR grant US4-DE019261, NIH/NCRR grant K30-RR022258, NIH/NIDCR grant U54-DE019261 and NIH/NHLBI grant R01-HL091005. Komro's and Maldonado-Molina's work was supported in part by NIH/NIAAA grants R01 AA013458, R01 AA016549, and K01 AA017480.

References

1. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.
2. Gurka MJ, Edwards LJ. Mixed models. *Handbook of Statistics, Volume 27*: *Epidemiology and Medical Statistics*. Elsevier, North-Holland: Amsterdam, 2008.
3. Box GEP, Draper NR. *Empirical Model-building and Response Surfaces*. Wiley: New York, 1987; 424.
4. Chatfield C. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A* 1995; **158**:419–466.
5. Draper D. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B* 1995; **57**:45–97.
6. Muller KE, Barton CN, Benignus VA. Recommendations for appropriate statistical practice in toxicologic experiments. *Neurotoxicology* 1984; **5**:113–126.
7. Muller KE, Fetterman BA. *Regression and ANOVA*: *An Integrated Approach Using SAS Software*. SAS Institute: Cary, NC, 2002.
8. Muller KE, Stewart PW. *Linear Model Theory; Univariate, Multivariate and Mixed Models*. Wiley: New York, 2006.
9. Kleinbaum DG, Kupper LL, Nizam A, Muller KE. *Applied Regression Analysis and Other Multivariable Methods* (4th edn). Duxbury Press: Boston, 2007.
10. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. Springer: New York, 2000.
11. Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Schabenberger O. *SAS for Mixed Models*. SAS Institute: Cary, NC, 2006.
12. Raudenbush SW. Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology* 2001; **52**:501–525.
13. Collins LM. Analysis of longitudinal data: the integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology* 2006; **57**:505–528.
14. Davidian M. *Applied Longitudinal Data Analysis*. Springer: New York, 2008.
15. Tuerlinckx F, Rijmen F, Verbeke G, De Boeck P. Statistical inference in generalized linear mixed models: a review. *British Journal of Mathematical and Statistical Psychology* 2006; **59**:225–255.
16. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal American Statistical Association* 1993; **88**:9–25.
17. Joiner BL. Lurking variables: some examples. *The American Statistician* 1981; **35**:227–233.
18. Edwards LJ, Muller KE, Wolfinger RD, Qaqish BF, Schabenberger O. An R^2 statistic for fixed effects in the linear mixed model. *Statistics in Medicine* 2008; **27**:6137–6157.
19. Welham SJ, Thompson R. Likelihood ratio tests for fixed model terms using residual maximum likelihood. *Journal of the Royal Statistical Society, Series B* 1997; **59**:701–714.
20. Shapiro A. Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika* 1985; **72**:133–144.
21. Shapiro A. Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review* 1988; **56**:49–62.
22. Stram DO, Lee JW. Variance components in the longitudinal mixed effects model. *Biometrics* 1994; **50**:1171–1177.
23. Stram DO, Lee JW. Correction to 'Variance components in the longitudinal mixed effects model'. *Biometrics* 1995; **51**:1196.

24. Atkinson AC. *Plots, Transformations, and Regression*. Clarendon Press: Oxford, 1985.
25. Schwarz G. Estimating the dimension of a model. *Annals of Statistics* 1978; **6**:461–464.
26. Keselman HJ, Algina J, Kowalchuk RK, Wolfinger RD. A comparison of two approaches for selecting covariance structures in the analysis of repeated measures. *Communications in Statistics*: *Simulation and Computation* 1998; **27**:591–604.
27. Burnham KP, Anderson DR. *Model Selection and Inference*: *A Practical Information-theoretic Approach*. Springer: New York, 1998.
28. Morrell CH, Brant LJ, Ferrucci L. Model choice can obscure results in longitudinal studies. *Journal of Gerontology*: *Medical Sciences A* 2009; **64**:215–222.
29. Wang J, Schaalje B. Model selection for linear mixed models using predictive criteria. *Communications in Statistics*: *Simulation and Computation* 2009; **38**:788–801.
30. Nilsson R, Bjorkegren J, Tegner J. Consistent feature selection for pattern recognition in polynomial time. *Journal of Machine Learning Research* 2007; **8**:589–612.
31. Pearl J. *Probablilistic Reasoning in Intelligent Systems*. Morgan Kauffman: San Fransisco, 1988.
32. Stinnett SS. Collinearity in mixed models. *Ph.D.*, University of North Carolina at Chapel Hill, 1993.
33. Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics*: *Identifying Influential Data and Sources of Collinearity*. Wiley: New York, 1980.
34. Gurka MJ, Edwards LJ, Muller KE, Kupper LL. Extending the Box–Cox transformation to the linear mixed model. *Journal of the Royal Statistical Society, Series A* 2006; **169**:255–272.
35. Komro KA, Perry CL, Veblen-Mortenson S, Bosma LM, Dudovitz BS, Williams CL, Jones-Webb R, Toomey TL. Brief report: the adaptation of Project Northland for urban youth. *Journal of Pediatric Psychology* 2004; **29**:457–466.
36. Komro KA, Perry CL, Veblen-Mortenson S, Farbakhsh K, Kugler KC, Alfano KA, Dudovitz BS, Williams CL, Jones-Webb R. Cross-cultural adaptation and evaluation of a home-based program for alcohol use prevention among urban youth: The 'Slick Tracy Home Team Program'. *Journal of Primary Prevention* 2006; **27**:135–154.
37. Komro KA, Maldonado-Molina MM, Tobler AL, Bonds JR, Muller KE. Effects of home access and availability of alcohol on young adolescents alcohol use. *Addiction* 2007; **102**:1597–1608.
38. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear model. *Biometrika* 1986; **73**:13–22.