Monte-Carlo Simulation in Modeling for Hierarchical Generalized Linear Mixed Models

Kyle M. Irimata and Jeffrey R. Wilson

Abstract It is common to encounter data that have a hierarchical or nested structure. Examples include patients within hospitals within cities, students within classes within schools, factories within industries within states, or families within neighborhoods within census tracts. These structures have become increasingly common in recent times and include variability at each level which must be taken into account. Hierarchical models which account for the variability at each level of the hierarchy, allow for the cluster effects at different levels to be analyzed within the models (Shahian et al. in Ann Thorac Surg, 72(6):2155–2168, 2001). This chapter discusses how the information from different levels can be used to produce a subject-specific model. However, there are often cases when these models do not fit as additional random intercepts and random slopes are added to the model. This addition of additional parameters often leads to non-convergence. We present a simulation study as we explore the cases in these hierarchical models which often lead to non-convergence. We also used the 2011 Bangladesh Demographic and Health Survey data as an illustration.

1 Introduction

Hierarchical logistic regression models consist of inherent correlation due to different sources of variation. At each level of the hierarchy, we have random intercepts and sometimes random slopes as well as the appropriate fixed effects. We have done extensive work with the GLIMMIX and NLMIXED procedures in fitting hierarchical models and have noted the trials and tribulations in computing regression estimates and covariance estimates associated with hierarchical models in SAS, as attested by

K.M. Irimata
School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ 85287, USA
e-mail: kirimata@asu.edu

J.R. Wilson (✉)
W.P. Carey School of Business, Arizona State University, Tempe, AZ 85287, USA
e-mail: Jeffrey.wilson@asu.edu

D.-G. Chen and J.D. Chen (eds.), *Monte-Carlo Simulation-Based Statistical Modeling*, ICSA Book Series in Statistics, DOI 10.1007/978-981-10-3307-0_13

others. We have had several occasions when our models do not converge. In some cases, we found that the convergence criterion was satisfied, but the standard error for the covariance parameters was given as "." This problem has gained the attention of many (Hartzel et al. 2001; Wilson and Lorenz 2015 to name a few). We do not know with certainty why certain convergence problems exist. As such we provide some understanding and make some suggestions based on our own work as well as work done by others. We also provide the steps and results of a simulation study which can be expanded upon for further exploration of the problem and its remedies.

In this chapter, we discuss the use of two-level and three-level hierarchical models for binary data, although it is possible to analyze higher level data. We discuss the use of models with effects at level 2 and level 3 representing random intercepts and random slopes. These random effects are added into the model to account for unobservable effects that are known to exist but were not measured or cannot be measured. We also discuss the use of simulations as a means of investigating issues or irregularities. This process is presented as an exercise in simulating hierarchical binary data, which for simplicity is restricted to the two-level case, although the techniques discussed can be readily expanded for higher levels. These simulated models have incorporated a random intercept and a random slope at level 2. We implement a hierarchical model using the GLIMMIX procedure in SAS, to identify factors that contribute to AIDS knowledge in Bangladesh and investigate models that do and do not converge based on the number of fixed effect predictors.

2 Generalized Linear Model

The birth of the generalized linear models unified many methods (Nelder and Wedderburn 1972). These models consist of a set of *n* independent random variables $Y_1 \ldots . Y_n$, each with a distribution from the exponential family. We define a generalized linear model as having three components: the random component, the systematic component, and the link component. We define the log-likelihood function based on unknown mean parameters, a dispersion parameter, and a weight parameter, denoted by θ_i, φ, and ω_i respectfully, and of the form (Smyth 1989),

$$l(\phi_i^{-1}, \omega_i : y_i) = \sum_i \{\omega_i \phi_i^{-1}[y_i\theta_i - b(\theta_i)] - c(y_i, \omega_i\phi_i^{-1})$$

with ϕ_i unknown and assume that

$$c(y_i, \omega_i\phi_i^{-1}) = \omega_i\phi_i^{-1}a(y_i) - \frac{1}{2}s(-\omega_i\phi_i^{-1}) + t(y_i)$$

Thus we have a generalized linear model for the mean such that

$$\mu_i = E(Y_i) = b'(\theta_i) = \mathbf{x_i'\beta}$$

where $\mathbf{x_i'} = (x_1, \ldots.., x_p)'$ is the vector of covariates and $\boldsymbol{\beta}$ is the vector of regression parameters. The functions a (y) and b (θ_i) are known functions. We also present the generalized linear model as

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

where the random component belongs to the exponential family of distributions, while in the marginal form we present $g(E(Y)) = \mathbf{X}\beta$. However, when the set of outcomes from the outcomes Y_i are not independent, then the generalized linear model in its pure form is no longer appropriate and we must use generalized linear mixed models.

3 Hierarchical Models

It is common in fields such as public health, education, demography, and sociology to encounter data structures where the information is collected based on a hierarchy. For instance, in health studies, we often see patients nested within doctors and doctors nested within hospitals. In these types of cases, there is variability at each level of the hierarchy, resulting in intraclass correlation due to the clustering. As a result of the correlation at each level inherent from these hierarchical structures, the standard logistic regression is inappropriate (Rasbash et al. 2012). Ignoring these levels of design while researching the outcome is sure to lead to erroneous results unless the intraclass correlation is of an insignificant size (Irimata and Wilson 2017). Others have demonstrated that ignoring a level of nesting in the data can impact variance estimates and the available power to detect significant covariates (Wilson and Lorenz 2015). When seeking to appropriately analyze these types of correlated data, we must extend the generalized linear models by accounting for the association among the responses.

Hierarchical models, also referred to as nested models or mixed models are statistical models that extend the class of generalized linear models (GLMs) to address and account for the hierarchical (correlated) nesting of data (Hox 2002; Raudenbush and Bryk 2002; Snijders and Bosker 1998). We will refer to these as the hierarchical generalized linear models (HGLMs). This approach incorporates a random effect, usually according to the normal distribution, although non-normal random effects can also be used. The extension required in HGLMs is not as involved when the outcomes follow a conditional normal distribution and the random effects are normally distributed. However, when dealing with outcomes that are not normally distributed (i.e. binary, categorical, ordinal), the extension is not as straightforward. In these cases, we often use a link other than the identity and must specify an appropriate

error distribution for the response at each level. We thus present the conditional mean explanation rather than the marginal mean.

While most work have concentrated on random intercepts, we have often been confronted with data requiring multiple random intercepts and even random slopes. When using the GLIMMIX procedure in SAS, we often find that models which include multiple random intercepts or even one random intercept with one random slope may not converge. Therefore, this chapter introduces the reader to hierarchical models with dichotomous outcomes (i.e., hierarchical generalized linear models), and provides concrete examples of non-convergence and possible remedies in these situations.

We present hierarchical models as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\theta + \varepsilon$$

where the random effects θ have a multivariate normal distribution with mean vector zero and covariance matrix G, with the distribution of the errors ε as normal with mean vector 0 and covariance matrix R. The $\mathbf{X}$ matrix consists of the fixed effects with vector of regression parameters β while the Z matrix consists of columns, each representing the random effects with vector of parameters θ. Researchers refer to this as compensating for the correlation through the systematic component. Thus we often write in the conditional response form as

$$g(E[\mathbf{Y}|\,\theta]) = \mathbf{X}\beta + \mathbf{Z}\theta$$

where $\theta \sim \mathcal{N}(0, G)$. The unconditional covariance matrix for Y, is

$$\text{var}(\mathbf{Y}) = \mathbf{A^{1/2}RA^{1/2}} + \boldsymbol{G}$$

and the conditional covariance matrix, given the random effects is given by

$$\text{var}(\mathbf{Y}|\theta) = \mathbf{A^{1/2}RA^{1/2}} = \mathbf{V}.$$

Thus, it is common in literature to refer to the G-side and R-side effects, which refer to the covariance matrix of the random effects, and the covariance matrix of the residual effects, respectfully.

In SAS, the GLIMMIX procedure distinguishes between the G-side and R-side effects and can model the random effects as well as correlated errors. This procedure fits generalized linear mixed models based on linearization and relies on a restricted pseudo-likelihood method of estimation. We revisit the method here as it helps us to understand the problems regarding non-convergence. This estimation is essentially based on the following.

Consider the conditional mean as

$$\mathrm{E}[\mathrm{Y}|\,\theta] = \mathrm{g}^{-1}(\mathbf{X}\beta + \mathbf{Z}\theta)$$

and using Taylor series expansion we linearize $\mathrm{g}^{-1}(\mathbf{X}\beta + \mathbf{Z}\theta)$ about the points $\tilde{\beta}$ and $\tilde{\theta}$ which gives

$$\mathrm{g}^{-1}(\mathbf{X}\beta + \mathbf{Z}\theta) \cong \mathrm{g}^{-1}\left(\mathbf{X}\tilde{\beta} + \mathbf{Z}\tilde{\theta}\right) + \frac{\partial \mathrm{g}^{-1}(\mathbf{X}\beta + \mathbf{Z}\theta)}{\partial \beta}\left(\beta - \tilde{\beta}\right) + \frac{\partial \mathrm{g}^{-1}(\mathbf{X}\beta + \mathbf{Z}\theta)}{\partial \theta}(\theta - \tilde{\theta})$$

$$\mathrm{g}^{-1}(\mathbf{X}\beta + \mathbf{Z}\theta) \cong \mathrm{g}^{-1}\left(\mathbf{X}\tilde{\beta} + \mathbf{Z}\tilde{\theta}\right) + \Omega_{|\tilde{\beta}\tilde{\theta}}\mathbf{X}\left(\beta - \tilde{\beta}\right) + \Omega_{|\tilde{\beta}\tilde{\theta}}\mathbf{Z}(\theta - \tilde{\theta})$$

where $\Omega_{|\tilde{\beta}}$ and $\Omega_{|\tilde{\theta}}$ denote the matrix of derivatives evaluated at $\tilde{\beta}$ and $\tilde{\theta}$ respectively. Thus

$$\Omega_{|\tilde{\beta}\tilde{\theta}}\{\mathrm{g}^{-1}(\mathbf{X}\beta + \mathbf{Z}\theta)\} \cong \Omega_{|\tilde{\beta}\tilde{\theta}}{}^{-1}\{\mathrm{g}^{-1}\left(\mathbf{X}\tilde{\beta} + \mathbf{Z}\tilde{\theta}\right)\} + \mathbf{X}\left(\beta - \tilde{\beta}\right) + \mathbf{Z}(\theta - \tilde{\theta})$$

So

$$\Omega_{|\tilde{\beta}\tilde{\theta}}\{\mathrm{g}^{-1}(\mathbf{X}\beta + \mathbf{Z}\theta)\} - \Omega_{|\tilde{\beta}\tilde{\theta}}{}^{-1}\{\mathrm{g}^{-1}\left(\mathbf{X}\tilde{\beta} + \mathbf{Z}\tilde{\theta}\right)\} \cong \mathbf{X}\beta + \mathbf{Z}\theta - (\mathbf{X}\tilde{\beta} + \mathbf{Z}\tilde{\theta})$$

and

$$\mathbf{X}\beta + \mathbf{Z}\theta \cong \left(\mathbf{X}\tilde{\beta} + \mathbf{Z}\tilde{\theta}\right) + \Omega_{|\tilde{\beta}\tilde{\theta}}{}^{-1}\{\mathrm{g}^{-1}(\mathbf{X}\beta + \mathbf{Z}\theta)\} - \Omega_{|\tilde{\beta}\tilde{\theta}}{}^{-1}\{\mathrm{g}^{-1}\left(\mathbf{X}\tilde{\beta} + \mathbf{Z}\tilde{\theta}\right)\}$$

$$\mathbf{X}\beta + \mathbf{Z}\theta \cong \left(\mathbf{X}\tilde{\beta} + \mathbf{Z}\theta\right) + \Omega_{|\tilde{\beta}\tilde{\theta}}{}^{-1}\{(\mathrm{E}[\mathbf{Y}|\,\theta]) - \{\mathrm{g}^{-1}\left(\mathbf{X}\tilde{\beta} + \mathbf{Z}\tilde{\theta}\right)]\}$$

Hence we consider the approximation and use the similar structure denoted by $\mathbf{X}\tilde{\beta} + \mathbf{Z}\tilde{\theta}$ to represent the matrix of fixed effects multiplied by a beta-like term and Z matrix of random effects multiplied by a theta-like term and we denote $\Omega_{|\tilde{\beta}\tilde{\theta}}{}^{-1}\left\{(\mathrm{E}[\mathbf{Y}|\,\theta]) - \{\mathrm{g}^{-1}\left(\mathbf{X}\tilde{\beta} + \mathbf{Z}\tilde{\theta}\right)\right\} = \zeta$ as an error-like term. So we can think of the approximation as a linear term and defined as

$$\mathrm{Y}_{\mathrm{approx}} = \mathbf{X}\beta + \mathbf{Z}\theta + \zeta$$

with the variance

$$\text{var}[\text{Y}_{\text{approx}}|\,\theta] = \text{var}[\{(\text{E}\,[\mathbf{Y}|\,\theta])] = \Omega_{|\tilde{\beta}\tilde{\theta}}{}^{-1}\mathbf{A^{1/2}RA^{1/2}}\Omega_{|\tilde{\beta}\tilde{\theta}}{}^{-1}$$

As such this can be seen as a linear approximation, given by Y_{approx} with fixed effects β, and random effects θ and variance of ζ given by $\text{var}[\text{Y}_{\text{approx}}|\,\theta]$.

3.1 Approaches with Binary Outcomes

Binary outcomes are very common in healthcare research, amongst many other fields. For example, one may investigate whether a patient has improved or recovered after discharge from the hospital or not. For healthcare and other types of research, the logistic regression model is one of the preferred methods of modeling data when the outcome variable is binary. In its standard form, it is a member of a class of generalized linear models specific to the binomial random component. As is customary in regression analysis, the model makes use of several predictor variables that may be either numerical or categorical. However, a standard logistic regression model assumes that the observations obtained from each unit are independent. If we were to fit a standard logistic regression to nested data, the assumption of independent observations is seriously violated. This violation could lead to an underestimation of the standard errors, which in turn can lead to conclusions of a significant effect, when in fact it is not.

Multilevel approaches for nested data can also be applied to analysis of dyadic data to take into account the nested sources of variability at each level (Raudenbush 1992). Many researchers have explored the use of these two-level approaches with binary outcomes (see for example McMahon et al. 2006).

4 Three-Level Hierarchical Models

In the analysis of multilevel data, each level provides a component of variance that measures intraclass correlation. For instance, consider a hierarchical model at three levels for the kth patient seeing the jth doctor in the ith hospital. The patients are at the lower level (level 1) and are nested within doctors (level 2) which are nested within hospitals at the next level (level 3). We consider the hospital as the primary unit, doctors as secondary unit, and patients as the observational unit. These clusters are treated as random effects. We make use of random effects as we believe there are some non-measurable influences on patient outcomes based on the doctor and also based on the hospital. Some effects may be positive and some effects may be negative, but overall we assume their average effects are zero.

4.1 With Random Intercepts

At level 1, we may take responses from different patients, while noting their age (Age) and length of stay (LOS). The outcomes are modeled through a logistic regression model

$$\log\left[\frac{p_{ijk}}{1-p_{ijk}}\right] = \gamma_{oij} + \gamma_{1ij}\text{Age}_{ijk} + \gamma_{2ij}\text{Los}_{ijk} \tag{4.1}$$

where γ_{oij} is the intercept, γ_{1ij} is the coefficient associated with the predictor Age_{ijk}, and γ_{2ij} is the coefficient associated with the predictor Los_{ijk} (length of stay) for $k = 1, 2, \ldots, n_{ij}$ patients; $j = 1, 2, \ldots, n_i$ doctors and $i = 1, \ldots, n$; hospitals. Each doctor has a separate logistic model. If we allow the effects of Age and LOS on the outcome to be the same for each doctor, but allow the intercept to be different on the logit scale, we have parallel planes for their predictive model. The γ_{oij} intercept represents those differential effects among doctors.

At level 2, we assume that the intercept γ_{oij} (which allows a different intercept for doctors within hospitals) depends on the unobserved factors specific to the ith hospital, the covariates given as associated with the doctors within the ith hospital, and a random effect u_{oij} associated with doctor j within hospital i. Thus,

$$\gamma_{oij} = \gamma_{oi} + \gamma_{1i}\text{Experience}_{ij} + u_{oij} \tag{4.2}$$

where Experience_{ij} is the experience for the jth doctor within the ith hospital. Similarly, hospital administration policies may have different effects on doctors. At level 3, the model assumes that differential hospital policies depend on the overall fixed intercept β_0 and the random intercept u_{oi} associated with the unmeasurable effect for hospital i. Thus,

$$\gamma_{oi} = \beta_0 + u_{oi} \tag{4.3}$$

By successive substitution into the expression for γ_{oi} in (4.3) into (4.2), and then by substituting the resulting expression for γ_{oij} into (4.1), we obtained

$$\log\left[\frac{p_{ijk}}{1-p_{ijk}}\right] = \beta_0 + \gamma_{1i}\text{Experience}_{ij} + \gamma_{1ij}\text{Age}_{ijk} + \gamma_{2ij}\text{Los}_{ijk} + u_{oi} + u_{oij} \tag{4.4}$$

The combination of random and fixed terms results in a generalized linear mixed model with two random effects; hospitals denoted by $u_{oi} \sim \mathcal{N}(0, \sigma^2_{u_i})$ and doctors denoted by $u_{oij} \sim \mathcal{N}(0, \sigma^2_{u_{ij}})$ with covariance $\sigma_{u_{oi}, u_{oij}}$. From Eq. (4.4), the model consists of the overall mean plus experience of doctors plus age of patient, length of stay plus effects due to hospitals and effects due to doctors for each individual. Hence, we have a subject-specific model.

4.2 Three-Level Logistic Regression Models with Random Intercepts and Random Slopes

Consider the three-level random intercept and random slope model consisting of a logistic regression model at level 1,

$$\log\left[\frac{p_{ijk}}{1-p_{ijk}}\right] = \gamma_{oij} + \gamma_{1ij}\text{Age}_{ijk} + \gamma_{2ij}\text{Los}_{ijk} \tag{4.5}$$

where both γ_{oij} and γ_{2ij} are random, for $k = 1, 2, \ldots, n_{ij}$; $j = 1, 2, \ldots, n_i$; and $i = 1, \ldots, n$. So each doctor has a different intercept and the rates of change with respect to length of stay are not the same for all the doctors. However, there are some unobserved effects related to LOS that impact remission. There are factors associated with LOS and the doctors' impacts on patients vary as LOS varies. The intercept represents a group of unidentifiable factors that impact the overall effect of the doctor on the patient's success, while the slope represents the differential impact that the particular variable (LOS) has that results in differences among patients.

So, at level 2, γ_{oij} and γ_{2ij} are treated as response variables within the model,

$$\gamma_{oij} = \gamma_{oi} + \gamma_{1i}\text{Experience}_{ij} + u_{oij} \tag{4.6}$$

$$\gamma_{2ij} = \gamma_{2i} + u_{2ij} \tag{4.7}$$

where γ_{oi} and γ_{2i} are random effects. Equation (4.6) assumes the intercept γ_{oij} for doctors nested within hospital j, depends on the unobserved intercept specific to the ith hospital, the effects associated with the doctor's experience in the hospital, and a random term u_{oij} associated with doctor j within hospital i. The slope γ_{2ij} depends on the overall slope γ_{2i} for hospital i and a random term u_{2ij}.

At level 3, the model shows that the hospitals vary based on random effects

$$\gamma_{oi} = \beta_{00} + u_{oi} \tag{4.8}$$

$$\gamma_{2i} = \beta_{22} + u_{2i} \tag{4.9}$$

The intercept γ_{oi} depends on the overall fixed intercept β_{00} and the random term u_{oi} associated with the hospital i, while the hospital slope γ_{2i} depends on the overall fixed slope β_{22} and the random effect u_{2i} associated with the slope for hospital i. By substituting the expression for γ_{oi} and γ_{2i} into (4.7) and (4.8), and then substituting the resulting expression for γ_{oij} and γ_{2ij} into (4.9), we obtained

$$\log\left[\frac{p_{ijk}}{1-p_{ijk}}\right] = \beta_{00} + \gamma_{1ij}\text{Age}_{ijk} + \gamma_{1i}\text{Experience}_{ij} + u_{oi} + u_{oij} + \left(\beta_{22} + u_{2i} + u_{2ij}\right)\text{Los}_{ijk} \tag{4.10}$$

Thus, we have a generalized linear mixed model with random effects u_{oi}, u_{oij}, γ_{1i} and γ_{1ij}. Therefore, Los_{ijk} is associated with both a fixed and random part. We take advantage of this regrouping of terms to incorporate the random effects and their variance-covariance matrix, so that u_{oi}, u_{oij}, γ_{1i} and γ_{1ij} are jointly distributed normally with a mean of zero and a covariance matrix reflecting the relationships between the random effects.

4.3 Nested Higher Level Logistic Regression Models

For higher than three level nested we can easily present the model, though executing the necessary computations may be tedious. Imagine if we had the data with another level, hospitals nested within cities (level 4 denoted by h). Cities may have their own way of monitoring healthcare within their jurisdiction. We also believed that the number of beds within the hospital is a necessary variable. For such data, we will have the kth patient nested within the jth doctor which is nested within ith hospital which is nested in the hth city. Then the model is:

$$\log\left[\frac{p_{hijk}}{1-p_{hijk}}\right] = \beta_{00} + \gamma_{1hij}\text{Age}_{hijk} + \gamma_{1hi}\text{Experience}_{hij} + \gamma_{1h}\text{Bed}_{hi} + u_{oh} + u_{ohi} + u_{ohij} + \left(\beta_{22} + u_{2hi} + u_{2hij}\right)\text{Los}_{hijk} \tag{4.11}$$

5 Possible Problems with Hierarchical Model

5.1 Issues in Hierarchical Modeling

We found that convergence of parameter estimates can sometimes be difficult to achieve, especially when fitting models with random slopes or higher levels of nesting. Some researchers have found that convergence problems may occur if the outcome is skewed for certain clusters or if there is quasi or complete separation. Such phenomena destroy the variability within clusters which is essential to obtaining the solutions. In addition, including too many random effects may not be computationally possible (Schabenberger 2005).

We also found what other researchers did; for hierarchical logistic models for nested binary data, it is often not feasible to estimate random effects for both intercepts and slopes at the same time in a model. Newsom (2002) showed that we can have models with too many parameters to be estimated given the number of covariance elements included. Others found that such models can lead to severe convergence problems, which can limit the modeling. Before fitting these conditional models, McMahon et al. (2006) suggested that one should determine whether there is

significant cluster interdependence to justify the use of multilevel modeling. Irimata and Wilson (2017) through simulation gave some further guidance.

Regardless of the number of clusters, Austin (2010) found that for all statistical software procedures, the estimation of variance components tended to be poor when there were only five subjects per cluster. The number of clusters on the mean number of quadrature points was negligible. However, when the random effects were large, Rodriquez and Goldman (1995) found substantial decreases in the estimation of fixed effects and/or variance components. They also found that there was bias in the estimation when the number of subjects per cluster was small.

These hierarchical models can be fitted through SAS with the GLIMMIX or NLMIXED procedure as well as in SPSS and R. Maas and Hox (2004) claimed that only one random statement is supported in the NLMIXED procedure so that nonlinear mixed models cannot be assessed at more than two levels. However, Hedeker et al. (2008), Hedeker et al. (2012) showed how more than one random statement can be used for continuous data in the NLMIXED procedure with more than two-levels.

5.2 Parameter Estimations

The conditional joint distribution of the responses and the distribution of the random effects provide a joint likelihood which cannot necessarily be readily written down in closed form. However, we still need to estimate the regression coefficients and the random components. In so doing, it is imperative for us to use some form of approximations. Sometimes researchers have used the quasi-likelihood approach through a Taylor series expansion to approximate the joint likelihood. The approximate likelihood is maximized to produce maximized quasi-likelihood estimates. The disadvantage which many researchers have pointed out with this approach is the bias involved with quasi-likelihoods (Wedderburn 1974). Other researchers have resorted to numerical integration, split up into quadratures, to obtain approximations of the true likelihood. More integration points will increase the number of computations and thus impede the speed to convergence, although it increases the accuracy. Each added random component increases the integral dimension. A random intercept is one dimension (one added parameter), a random slope makes that two dimensions. Our experience is that the three-level nested models with random intercepts and slopes often create problems regarding convergence.

5.3 Convergence Issues in SAS

We spent considerable time overcoming the challenges of the GLIMMIX procedure. We reviewed available literature and discussed with those with experience using SAS. Although there are by no means guarantees that there will not be challenges, we provide in this chapter our experiences, underscored by others, as well as suggestions for improving the performance of this procedure.

Non-convergence in the GLIMMIX procedure can be identified by looking at the output and the log. The most obvious indication of issues is in the convergence criterion, which is provided below the iteration history. When convergence is not obtained, SAS will provide the following warning: "DID NOT CONVERGE".

A successful convergence message does not itself necessarily guarantee that the model converged. In some cases, the convergence criterion will be satisfied, but the standard error for one or more of the (non-zero) covariance parameters will be missing. When this occurs, the standard error will be given by a "." instead of an actual estimate. In these cases, the output may look similar to the following:

Covariance Parameter Estimates

Cov Parm	*Subject*	*Estimate*	*Standard Error*
Intercept	div	0.09097	.
urban	div	0.01127	.

When there is non-convergence, there are a number of possible remedies. Many authors, such as Kiernan et al. (2012) have offered a number of possible solutions. Researchers using the GLIMMIX procedure may choose to:

- Drop certain variables
- Relax the convergence criterion
- Increase the value of ABSCONV =
- Change the covariance structure using TYPE =
- Adjust the quadrature using QUAD =
- Utilize different approximation algorithms such as TECH = NRRIDG or TECH = NEWRAP, in the NLOPTIONS statement.
- Increase the number of iterations using MAXITER = in the NLOPTIONS statement
- Control the number of outer iterations using the INITGLM option
- Increase the number of optimizations using the MAXOPT = option
- Rescale data values to reduce issues relating to extreme values
- Utilize an alternative approach, such as the %HPGLIMMIX MACRO (Xie and Madden 2014)

For a more thorough discussion of the procedure itself, Ene et al. (2015) provided a thorough introduction to the use and interpretation of the GLIMMIX procedure in SAS.

6 Simulation of Data

The IML procedure in SAS was used to simulate two-level data following a generalized linear mixed model with random intercepts and random slopes. In this example,

we explored the effects of including an increasing number of fixed effects when using the GLIMMIX procedure to fit a logistic regression model with one random intercept and one random slope. The approaches discussed in this section can readily be expanded to simulate data with more than two levels, although only two levels are discussed for ease of interpretation and understanding.

6.1 Simulation Setup

Here we set the parameters for the simulation. We will assume that our random intercept has variance σ^2_{INT}=7 and that the random slope has variance σ^2_{SLOPE}=15. We also assume that there are six continuous fixed effects. Each of the fixed effects has a mean of 1, with some random noise added such that the means are not all equal. The fixed effects are assumed to independent of one another and also pairwise independent of the random slope. The simulated data will include 15 clusters of observations, each with a randomly chosen number of observations between 2 and 40.

```
proc iml;
*Set the variance of the random slope;
sigInt = 7;
sigSlope = 15;
*Set the coefficients;
Bcont = 0.09;
*Set the observation level parameters;
*Set the means for 6 continuous fixed effects, and one
random slope;
means = {1,1,1,1,1,1,6};
*Slightly alter the means;
noise = normal(j(7,1));
noise[7]=0;
means = means+noise;
*Set the covariance of the fixed and random predictors;
R=I(7);
*Select the number of clusters;
b = 15;
*Randomly select the number of observations in each clus-
ter;
randobs = j(b,1);
call randgen(randobs, "Uniform");
*Transform to be between 2 and 40;
n = 2 + floor((41-2)*randobs);
*Calculate the overall total number of observations;
ntot = sum(n);
```

Once the parameters for the simulation are chosen, the cluster level data are created. Each of the random (cluster) intercepts are chosen according to independent random standard normal distributions with mean of 0 and standard deviation of 1.

The random (cluster) slope coefficients are also chosen according to independent random normal distributions with our specified variance and a mean of -1. In effect, each of the 15 clusters is assigned a unique cluster level intercept and slope term. Our design matrix is created using these random values.

```
*Create cluster level data;
*Cluster IDs;
cid = (1:b)`;
*Cluster random intercepts;
cint = randnormal(b,0,1);
*Cluster random slopes;
cslope = randnormal(b,-1,sigSlope);
*Loop through to create the design matrix;
cluster = j(ntot, 3);
startindex = 1;
do i=1 to b;
       endindex = startindex + n[i] - 1;
       cluster[startindex:endindex,1] = cid[i];
       cluster[startindex:endindex,2] = cint[i];
       cluster[startindex:endindex,3] = cslope[i];
       startindex = endindex + 1;
end;
```

Once the cluster level data are created, we can generate the observation level data. We create a matrix of independent normal realizations to serve as the observations for each of the six continuous fixed as well as the random slope variables. The realizations of each variable are created using a multivariate random normal. The fixed effect predictors are also transformed for better model fitting.

```
*Create observation level data;
X = randnormal(ntot, means, R);
*Apply some changes to the observation level data;
X[,1:6] = (X[,1:6]/1.6 + 5.1)*10;2] = bin(X[,2],cuts) - 1;
```

We combine our simulated data to create two matrices. The first matrix is used to combine all fixed and random effects information, while the second matrix provides a reduced set of information for use in simulation of the response. This second matrix removes information on the true random slope coefficient and the true cluster ID and thus contains information on the six fixed effects and the random intercept term.

```
*Create matrix of both cluster and observation level data;
alldat = X || cluster;
*Final data for simulation, excluding the random slope
predictor and cluster ID;
keepind = {1,2,3,4,5,6,9};
simdat = alldat[,keepind];
```

The coefficients for the fixed effect predictors are set according to those specified at the start of the simulation. The cluster level (random) intercept is assigned a coefficient equal to the square root of the random intercept variance term; since the random intercepts were originally simulated from a standard normal distribution, this coefficient introduces the specified variance into the simulation. These coefficients are also standardized based on the standard deviation of the respective observations.

```
*Set coefficients;
beta = j(7,1);
beta[1:6]=Bcont;
beta[7] = sqrt(sigInt);
*Standardize betas by the standard deviation;
datadev = STD(simdat);
beta = beta / datadev`;
```

We create our response as a function of these covariates. The simulated data are multiplied by the coefficients and the effect of the random slope is added in. The resulting value is then converted into a probability and used to create a binary response according to the Bernoulli distribution. This response is then combined with a "blinded" data matrix which has the value of the cluster intercept and the random slope coefficients removed. The final matrix is then output to a SAS data set with specified variable names.

```
*Create the response with the random slope effect added;
xb = simdat * beta + cluster[,3] # alldat[,7];
probs = 1 / (1 + exp(-xb));
y = rand("Bernoulli",probs);
*Create the final data with the cluster intercept removed;
outdat = y || alldat[,1:8];
*Output to a data set;
create SimData from
       outdat[colname={"Y" "X1" "X2" "X3" "X4" "X5" "X6"
"Xclust" "CID"}];
append from outdat;
close SimData;
*Quit IML;
quit;
```

The outputted data set is then analyzed using the GLIMMIX procedure in SAS. Each of the fixed effect predictors is added to the model one by one to determine the point at which this procedure will fail, if at all. A partial example of these analyses are shown below.

```
*Analyze the data using glimmix;
*One fixed effect;
proc glimmix data=SimData1;
       class CID;
       model Y(event="1") = X1 Xclust / dist=binary
link=logit;
       random intercept Xclust / type=vc subject=CID;
run;

*Two fixed effects;
proc glimmix data=SimData;
       class CID;
       model Y(event="1") = X1 X2 Xclust / dist=binary
link=logit;
       random intercept Xclust / type=vc subject=CID;
run;

[…]

*Six fixed effects;
proc glimmix data=SimData;
       class CID;
       model Y(event="1") = X1 X2 X3 X4 X5 X6 Xclust /
dist=binary link=logit;
       random intercept Xclust / type=vc subject=CID;
run;
```

6.2 Simulation Results

Although the GLIMMIX procedure is a powerful tool for fitting generalized linear models, it is not uncommon to find that the procedure fails to provide results. We utilized a simulation study similar to the one utilized in the previous section to investigate the effect of the number of predictors on the failure rates in the GLIMMIX procedure. A SAS macro was implemented to run the simulation across a variety of conditions and the GLIMMIX procedure was used to analyze the data under each condition for 1000 replications per condition. Each simulated data set contained information on a binary outcome, an identifier label for cluster number, one (random) cluster level predictor, and six fixed effect predictors. For each simulated data set, the GLIMMIX procedure was used to analyze the data set six times, where each call to the procedure included one additional fixed effect predictor.

Table 1 Failure rates for the GLIMMIX procedure (three clusters)

	Variances		Number of predictors					
Beta	Intercept	Slope	1	2	3	4	5	6
Weak	Low	Low	1	0.773	0.783	0.786	0.799	0.844
Moderate	Low	Low	1	0.760	0.784	0.804	0.798	0.815
Strong	Low	Low	1	0.780	0.785	0.807	0.817	0.821
Weak	Low	Medium	0	0.418	0.638	0.743	0.858	0.941
Moderate	Low	Medium	0	0.501	0.706	0.856	0.903	0.934
Strong	Low	Medium	0	0.325	0.506	0.675	0.788	0.893
Weak	Low	High	0	0.423	0.577	0.684	0.793	0.877
Moderate	Low	High	0	0.549	0.689	0.820	0.898	0.911
Strong	Low	High	0	0.450	0.629	0.703	0.826	0.865
Weak	Medium	Low	0	0.492	0.644	0.751	0.823	0.912
Moderate	Medium	Low	0	0.399	0.518	0.675	0.811	0.885
Strong	Medium	Low	0	0.322	0.488	0.641	0.745	0.817
Weak	Medium	Medium	0	0.459	0.604	0.698	0.820	0.899
Moderate	Medium	Medium	0	0.423	0.607	0.760	0.856	0.923
Strong	Medium	Medium	0	0.428	0.537	0.648	0.761	0.846
Weak	Medium	High	0	0.422	0.610	0.716	0.844	0.902
Moderate	Medium	High	0	0.367	0.557	0.725	0.846	0.910
Strong	Medium	High	0	0.393	0.543	0.636	0.788	0.831
Weak	High	Low	0	0.463	0.565	0.712	0.791	0.877
Moderate	High	Low	0	0.392	0.515	0.662	0.845	0.885
Strong	High	Low	0	0.364	0.509	0.664	0.743	0.851
Weak	High	Medium	0	0.529	0.701	0.777	0.880	0.956
Moderate	High	Medium	0	0.413	0.602	0.684	0.854	0.915
Strong	High	Medium	0	0.356	0.511	0.669	0.769	0.845
Weak	High	High	0	0.324	0.519	0.661	0.779	0.839
Moderate	High	High	0	0.327	0.484	0.656	0.831	0.869
Strong	High	High	1	0.376	0.581	0.738	0.808	0.842

In particular, the conditions examined were the number of data clusters, the strength of the variance of both the random intercept and slope and strength of fixed effect coefficients. The simulation took into account data sets with either 3, 15 or 45 clusters of data. The random effect variances we investigated included all combinations of low, medium and high variances for the random intercept and random slope—yielding a total of nine different variance combinations. The fixed effects also took three levels of strength—weak, moderate or strong.

The results of this simulation are given in Tables 1, 2 and 3 and are also displayed graphically in Fig. 1. These displays provide the failure rates for the 1000 simulations conducted for each of the specified conditions, thus higher values indicate poorer

Table 2 Failure Rates for the GLIMMIX Procedure (fifteen clusters)

	Variances		Number of predictors					
Beta	Intercept	Slope	1	2	3	4	5	6
Weak	Low	Low	0	0.413	0.400	0.404	0.414	0.396
Moderate	Low	Low	0	0.415	0.424	0.433	0.434	0.422
Strong	Low	Low	0	0.441	0.430	0.431	0.418	0.428
Weak	Low	Medium	0	0.138	0.180	0.337	0.484	0.610
Moderate	Low	Medium	0	0.182	0.272	0.387	0.493	0.579
Strong	Low	Medium	0	0.121	0.199	0.273	0.378	0.459
Weak	Low	High	0	0.152	0.281	0.401	0.509	0.635
Moderate	Low	High	0	0.200	0.269	0.355	0.459	0.545
Strong	Low	High	0	0.131	0.239	0.332	0.370	0.473
Weak	Medium	Low	0	0.171	0.251	0.321	0.449	0.567
Moderate	Medium	Low	0	0.148	0.258	0.445	0.570	0.613
Strong	Medium	Low	0	0.093	0.157	0.222	0.328	0.411
Weak	Medium	Medium	0	0.148	0.243	0.325	0.411	0.530
Moderate	Medium	Medium	0	0.167	0.294	0.406	0.514	0.573
Strong	Medium	Medium	0	0.118	0.189	0.284	0.348	0.455
Weak	Medium	High	0	0.214	0.304	0.396	0.459	0.627
Moderate	Medium	High	0	0.220	0.295	0.399	0.478	0.590
Strong	Medium	High	0	0.158	0.238	0.305	0.404	0.489
Weak	High	Low	0	0.092	0.191	0.324	0.404	0.531
Moderate	High	Low	0	0.157	0.252	0.366	0.440	0.552
Strong	High	Low	0	0.085	0.141	0.249	0.351	0.446
Weak	High	Medium	0	0.129	0.217	0.347	0.446	0.532
Moderate	High	Medium	0	0.122	0.227	0.335	0.503	0.605
Strong	High	Medium	0	0.133	0.194	0.272	0.359	0.437
Weak	High	High	0	0.140	0.232	0.329	0.470	0.541
Moderate	High	High	0	0.093	0.173	0.258	0.369	0.505
Strong	High	High	0	0.129	0.199	0.266	0.332	0.465

performance as a higher proportion of the calls to the GLIMMIX procedure failed to provide results. Tables 1, 2 and 3 divide the results of the simulations based on the number of clusters in each simulation, where Table 1 summarizes the simulations with 3 data clusters each, Table 2 summarizes the simulations with 15 data clusters each and Table 3 summarizes the simulations with 45 data clusters each. The first column in each of the tables provides the strength of the fixed effects predictor (weak, moderate or strong). The second and third columns denote the simulation settings for the variance of the random intercept and slope, respectively, where each variance term

Table 3 Failure Rates for the GLIMMIX Procedure (forty-five clusters)

	Variances		Number of predictors					
Beta	Intercept	Slope	1	2	3	4	5	6
Weak	Low	Low	0	0.243	0.247	0.245	0.236	0.235
Moderate	Low	Low	1	0.236	0.246	0.249	0.255	0.244
Strong	Low	Low	0	0.298	0.303	0.305	0.296	0.303
Weak	Low	Medium	0	0.055	0.090	0.123	0.159	0.184
Moderate	Low	Medium	0	0.051	0.098	0.127	0.163	0.243
Strong	Low	Medium	1	0.038	0.073	0.108	0.141	0.189
Weak	Low	High	1	0.035	0.073	0.099	0.140	0.197
Moderate	Low	High	1	0.038	0.061	0.101	0.157	0.214
Strong	Low	High	0	0.041	0.052	0.081	0.139	0.197
Weak	Medium	Low	0	0.060	0.089	0.134	0.169	0.223
Moderate	Medium	Low	0	0.061	0.093	0.127	0.189	0.215
Strong	Medium	Low	0	0.045	0.069	0.096	0.134	0.185
Weak	Medium	Medium	0	0.027	0.054	0.088	0.163	0.209
Moderate	Medium	Medium	0	0.062	0.104	0.139	0.168	0.208
Strong	Medium	Medium	0	0.071	0.099	0.118	0.146	0.187
Weak	Medium	High	0	0.041	0.088	0.130	0.177	0.209
Moderate	Medium	High	0	0.050	0.066	0.112	0.143	0.169
Strong	Medium	High	0	0.026	0.052	0.089	0.126	0.169
Weak	High	Low	0	0.031	0.057	0.080	0.141	0.209
Moderate	High	Low	0	0.031	0.061	0.104	0.139	0.187
Strong	High	Low	0	0.032	0.046	0.092	0.156	0.192
Weak	High	Medium	0	0.050	0.094	0.139	0.206	0.263
Moderate	High	Medium	0	0.068	0.111	0.161	0.182	0.259
Strong	High	Medium	0	0.030	0.068	0.101	0.129	0.173
Weak	High	High	0	0.038	0.086	0.139	0.170	0.235
Moderate	High	High	0	0.081	0.123	0.179	0.244	0.300
Strong	High	High	0	0.053	0.080	0.125	0.149	0.202

takes one of three levels (low, medium, high). The remaining six columns contain the failure rates as a proportion for the GLIMMIX procedure with a given number of fixed effects predictors. For instance, we can see from Table 1, in the first data row, in the last column that of the 1000 simulations with three clusters, weak fixed effects, low intercept variance and low slope variance, 84.4 % of the models with six fixed effect predictors failed to converge.

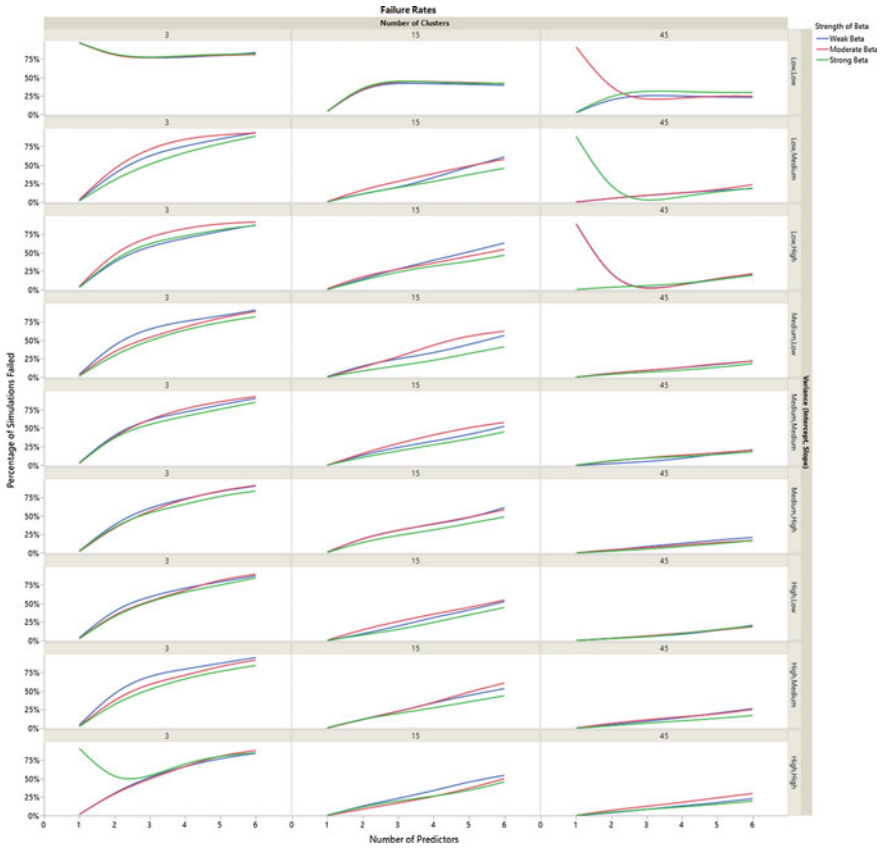

Fig. 1 Failure rates for the GLIMMIX procedure

Figure 1 provides a graphical representation of the same simulated data presented in Tables 1, 2 and 3. Each individual plot contains three lines representing the failure rates for each of the three strengths of the fixed effects. The blue line represents the simulations with weak predictors, the red line represents the simulations with moderate predictors and the green line represents the simulations with strong predictors. The vertical (Y) axis of each individual plot denotes the failure rates as a percentage, where higher values indicate higher rates of failure. The horizontal (X) axis within each of the individual plots represents the number of fixed effects included in the model for those simulations. The individual plots are also organized into three columns according to the number of data clusters in those simulations. The individual plots are further grouped into nine rows according to the strength of the random effects for those simulations. For example, in the individual plot in the last column of the first row contains information on the 1000 simulations in which there were 45 clusters, with weak fixed effects predictors, low random intercept variance and low random slope variance.

In general, as the number of predictors increased, the failure rates also increased. Notable exceptions include the case where there is very little variance in the random effects. For instance, in the case of low random intercept variance and low random slope variance, the failure rates may actually decrease, or increase only slightly. We can also see that the effect of increasing the number of predictors is also suppressed when there are more data clusters. In general, the GLIMMIX procedure is more successful in analyzing data with more clusters as illustrated by the lower failure rates. Similarly, data with overall stronger random effect variance is also less susceptible to failure as the number of predictors in the model increases. This holds true with respect to both the random intercept variance as well as the random slope variance.

7 Analysis of Data

7.1 Description

A subset of data from the 2011 Bangladesh Demographic and Health Survey is used in this study. This subset contains information on 1000 women between the ages of 10 and 49, living in Bangladesh. The data in this study are hierarchical in nature in that each of the women is nested within one of seven different districts, which correspond approximately to administrative regions in Bangladesh (NIPORT 2013). A simplified version of this structure is represented as Fig. 2.

The outcome of interest in this data set is a binary variable representing the woman's knowledge of AIDS. The variable takes one of two values representing knowledge of AIDS (1) or no knowledge of AIDS (0). In addition to this outcome, the data set also includes information on the woman's wealth index, age, number of living children as well as whether or not the woman lives in an urban or rural setting. Wealth index had five possible levels representing the quintile to which the woman belonged. Age represented the woman's age at the time of survey while number of living children represented how many living children the woman had at the time of survey. The urban/rural variable was a district level predictor as the value of this predictor were partially driven by the administrative region.

Please note to use the included DHS subset data, you must register as a DHS data user at: http://www.dhsprogram.com/data/new-user-registration.cfm. This subset data must not be passed on to others without the written consent of DHS (archive@dhsprogram.com). You are required to submit a copy of any reports/publications resulting from using this subset data to: archive@dhsprogram.com.

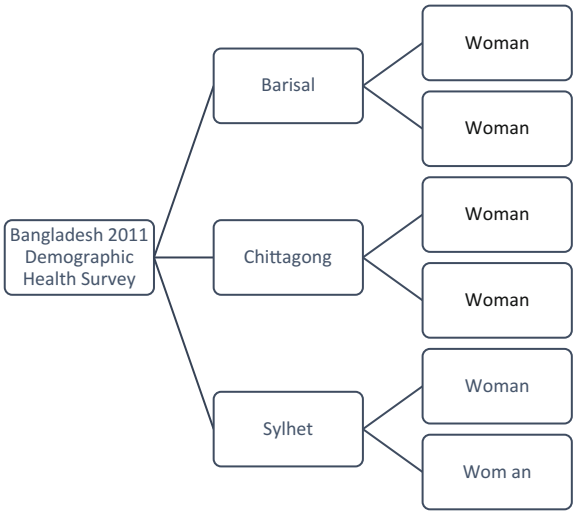

Fig. 2 Hierarchical structure in 2011 DHS Study

7.2 Data Analysis

We fit a logistic regression model with one random intercept and one random slope for the urban/rural variable. For these data, the random effects were used to address the clustering present due to districts. Each of these models was fitted using the GLIMMIX procedure in SAS. The first model included one fixed effect predictor for wealth index.

$$\log\left[\frac{p_{jk}}{1-p_{jk}}\right] = \beta_0 + \gamma_1 \text{Urban}_j + \gamma_{1j}\text{Wealth}_{jk} + u_{oj}$$

As in the data simulation section, these data can be analyzed in SAS using code similar to the example given below. Note that additional fixed effects predictors can be included in the model statement to fit additional models.

```
proc glimmix data=bang;
       class div urban wealth;
       model aids(event="1") = urban wealth / dist=binary
link=logit;
       random intercept urban /type=vc subject=div;
run;
```

The convergence criterion noted that the GLIMMIX procedure converged successfully and that we are also provided with standard errors for our random effects. Therefore, we see the procedure was successful in fitting the model.

Iteration History					
Iteration	*Restarts*	*Subiterations*	*Objective Function*	*Change*	*Max Gradient*
0	*0*	*4*	4562.1081534	2.00000000	0.00012
1	*0*	*3*	4679.3151727	0.37988319	0.000023
2	*0*	*2*	4718.2025244	0.05445086	0.000019
3	*0*	*1*	4720.2027844	0.00248043	0.000042
4	*0*	*1*	4720.2171672	0.00004268	1.244E-8
5	*0*	*0*	4720.2172745	0.00000000	5.905E-6

Convergence criterion (PCONV=1.11022E-8) satisfied.

Fit Statistics	
-2 Res Log Pseudo-Likelihood	4720.22
Generalized Chi-Square	973.41
Gener. Chi-Square / DF	0.98

Covariance Parameter Estimates			
Cov Parm	*Subject*	*Estimate*	*Standard Error*
Intercept	div	0.1149	0.1102
urban	div	0.05142	0.09565

Type III Tests of Fixed Effects				
Effect	*Num DF*	*Den DF*	*F Value*	*Pr > F*
urban	1	6	1.48	0.2692
wealth	4	982	21.92	<.0001

We also fit the model which included fixed effects for both wealth and age.

$$\log\left[\frac{p_{jk}}{1-p_{jk}}\right] = \beta_0 + \gamma_1 \mathrm{Urban}_j + \gamma_{1j} Wealth_{jk} + \gamma_{2j} \mathrm{Age}_{jk} + u_{oj}$$

In this case, we can similarly see that the convergence criterion is satisfied and that estimates of the standard errors of the random effects are provided. Thus, we see that the GLIMMIX procedure was successful in fitting a model.

Iteration History

Iteration	*Restarts*	*Subiterations*	*Objective Function*	*Change*	*Max Gradient*
0	*0*	*4*	4565.1551673	2.00000000	0.000164
1	*0*	*3*	4759.8377085	0.81042673	0.00017
2	*0*	*2*	4808.4720043	0.11542518	0.000114
3	*0*	*1*	4811.3231643	0.00580829	0.000136
4	*0*	*1*	4811.3435049	0.00011603	5.451E-8
5	*0*	*1*	4811.3434869	0.00000132	5.917E-9
6	*0*	*0*	4811.3434867	0.00000000	1.381E-7

Convergence criterion (PCONV=1.11022E-8) satisfied.

Fit Statistics

-2 Res Log Pseudo-Likelihood	4811.34
Generalized Chi-Square	973.16
Gener. Chi-Square / DF	0.98

Covariance Parameter Estimates

Cov Parm	*Subject*	*Estimate*	*Standard Error*
Intercept	div	0.1152	0.1059
urban	div	0.03594	0.08961

Type III Tests of Fixed Effects

Effect	*Num DF*	*Den DF*	*F Value*	*Pr > F*
urban	1	6	1.70	0.2403
wealth	4	981	22.06	<.0001
age	1	981	43.47	<.0001

We added a third predictor for *number of living children* to our mixed model.

$$\log\left[\frac{p_{jk}}{1-p_{jk}}\right] = \beta_0 + \gamma_1 \mathrm{Urban}_j + \gamma_{1j}\mathrm{Wealth}_{jk} + \gamma_{2j}\mathrm{Age}_{jk} + \gamma_{3j}\mathrm{Children}_{jk} + u_{oj}$$

With the inclusion of this third predictor, we see that the GLIMMIX procedure fails to converge and consequently does not provide estimates of the standard errors of the random effects. Hence, we see that, although SAS is able to fit the model with two fixed effects, the inclusion of a third fixed effect leads to failure.

Iteration History

Iteration	*Restarts*	*Subiterations*	*Objective Function*	*Change*	*Max Gradient*
0	*0*	*4*	4583.476774	2.00000000	3.121596
1	*0*	*3*	4788.3397625	2.00000000	6.129E-6
2	*0*	*2*	4842.5826312	0.36041184	0.000157
3	*0*	*1*	4846.0866446	0.19708062	0.000187
4	*0*	*1*	4847.0951336	0.16428366	1.49E-7
5	*0*	*1*	4848.0956362	0.14098626	2.589E-9
6	*0*	*1*	4849.0959182	0.12352732	7.98E-9
7	*0*	*1*	4850.096021	0.10993349	4.498E-8
8	*0*	*1*	4851.0960586	0.09904095	6.01E-11
9	*0*	*0*	4852.0960724	0.09011439	4.282E-6
10	*0*	*0*	4853.0960785	0.08266460	5.821E-6
11	*0*	*0*	4854.0960808	0.07635276	6.385E-6
12	*0*	*0*	4855.0960816	0.07093651	6.594E-6
13	*0*	*0*	4856.0960819	0.06623786	6.674E-6
14	*0*	*0*	4857.0960819	0.06212302	6.687E-6
15	*0*	*0*	4858.0960828	0.05848863	6.683E-6
16	*0*	*0*	4859.0960811	0.05525857	6.782E-6
17	*0*	*0*	4860.0960789	0.05236722	6.898E-6
18	*0*	*0*	4861.0960963	0.04974318	5.95E-6
19	*0*	*0*	4862.0961081	0.04737287	7.581E-6

Did not converge.

Covariance Parameter Estimates

Cov Parm	*Subject*	*Estimate*	*Standard Error*
Intercept	div	0.09097	.
urban	div	0.01127	.

Although we do not explore its use in depth here, the %hpglimmix macro provides an alternative approach in SAS for fitting generalized linear mixed models (Xie and Madden 2014). This macro offers improvements in memory usage as well as processing time and supports the fitting of more complicated models as compared to the GLIMMIX procedure. Although this macro does not currently provide standard errors of the covariance parameter estimates or Type III test results, it can be useful when alternative approaches fail to resolve convergence issues in the GLIMMIX procedure. We fit the previously discussed model, which includes three fixed effects predictors as well as one random intercept and one random slope for the Bangladesh data. After loading the macro into the current SAS session, the model can be run using code similar to the following.

```
%hpglimmix(data=bang,
      stmts=%str(
            class div urban wealth children;
            model aids = urban wealth age children / solu-
tion ;
            random int urban / subject=div solution;
      ),
      error=binomial, maxit=50,
      link=logit
);
```

Though this model fails to converge in the GLIMMIX procedure, we see that %HPGLIMMIX provides results for the model which includes three fixed effect predictors.

Iteration History

Iteration	*Evaluations*	*Objective Function*	*Change*	*Max Gradient*
0	4	4892.189763	.	6.524357
1	5	4892.1622318	0.02753122	5.886723
2	3	4892.1566979	0.00553391	5.959108
3	3	4892.1564908	0.00020710	5.957165
4	5	4892.0850182	0.07147255	3.174453
5	4	4892.0847449	0.00027336	3.145538
6	4	4892.0726771	0.01206780	0.563614
7	4	4892.0726558	0.00002129	0.569235
8	4	4892.0726553	0.00000047	0.568549
9	5	4892.0724006	0.00025469	0.41386
10	*4*	4892.0721478	0.00025279	0.01987

Convergence criterion (GCONV=1E-8) satisfied.

Covariance Parameter Estimates

Cov Parm	*Subject*	*Estimate*
Intercept	div	0.09138
urban	div	0.01301
Residual		0.9905

Fit Statistics

-2 Res Log Likelihood	4892.07215
AIC (smaller is better)	4898.07215
AICC (smaller is better)	4898.09666
BIC (smaller is better)	4897.90988
CAIC (smaller is better)	4900.90988

Another possible remedy in this case is found in the NLMIXED procedure in SAS. This procedure utilizes likelihood-based approaches to fit mixed models for nonlinear outcomes (Wolfinger 1999). This procedure is readily available in SAS software and provides similar techniques to those available in the GLIMMIX procedure. Although the models that can be fit in both procedures are similar, it is worth noting that the two procedures use different techniques for estimation and thus the results may vary between the two approaches. However, because different estimation techniques are employed there are also cases in which the NLMIXED procedure will converge, while the GLIMMIX procedure will not.

The NLMIXED procedure is implemented differently as compared to many other procedures in SAS software. In particular, one must provide starting values for each of the parameters of interest, which can be estimated in a number of ways. In this example, we first used the logistic procedure to obtain estimates of the fixed effects parameters and specify a generic value of '1' for the variance of each of our random effects (intercept and slope). We also specified an equation with respect to our parameters and observed predictor values, and use this equation in the specification of our model statement through the calculation of our probability using the logit link. Finally, each of the random effects as well as the corresponding distribution is specified, and the subject assigned.

```
proc logistic data=bang;
       model aids_knowledge(event="1") = urban wealth age
children/ link=logit;
run;

proc nlmixed data=bang;
       parms b0=0.6916 b1=0.3335 b2=0.5921 b3=-0.0287 b4=-
0.2616 s2u = 1 s2r = 1;
       xb = b0 + u + (b1+rb1)*urban + b2*wealth + b3*age +
b4*children;
       p = exp(xb) / (1+exp(xb));
       model aids_knowledge ~binary(p);
       random u rb1 ~ normal([0,0],[s2u,0,s2r]) sub-
ject=div;
run;
```

We found that the NLMIXED procedure converges successfully and also provides solutions for both our fixed and random effects for the model which includes three fixed effects predictors. Wolfinger (1999) provides a good introduction to the NLMIXED procedure and its usage, as well as some of the underlying calculations.

NOTE: GCONV convergence criterion satisfied.

Fit Statistics

-2 Log Likelihood	972.1
AIC (smaller is better)	986.1
AICC (smaller is better)	986.2
BIC (smaller is better)	985.7

Parameter Estimates

Parameter	*Estimate*	*Standard Error*	*DF*	*t Value*	*Pr > |t|*	*Alpha*	*Lower*	*Upper*	*Gradient*
b0	0.6547	0.3339	5	1.96	0.1072	0.05	–0.204	1.5131	–0.0001
b1	0.3267	0.2801	5	1.17	0.2960	0.05	–0.393	1.0468	0.00032
b2	0.6156	0.06906	5	8.91	0.0003	0.05	0.4381	0.7931	–0.0004
b3	–0.0296	0.01113	5	–2.66	0.0448	0.05	–0.058	–0.001	0.00999
b4	–0.2567	0.06031	5	–4.26	0.0080	0.05	–0.412	–0.102	0.0018
s2u	0.03178	0.04894	5	0.65	0.5450	0.05	– 0.094	0.1576	0.00059
s2r	0.2798	0.2937	5	0.95	0.3846	0.05	–0.475	1.0349	–0.0001

In general, we found that the results of our data analysis are in agreement with our findings based on the simulation study. The GLIMMIX procedure was successful in analyzing the models with fewer fixed effects predictors. However, once we included additional fixed effects, we saw that the GLIMMIX procedure failed to converge. In these cases, we may choose to investigate only the smaller subset of predictors in order to get successful analyses. Alternatively, if the larger number of predictors is of interest, we can utilize the %HPGLIMMIX macro, which is able to achieve convergence, although the output is reduced. We may also utilize the NLMIXED procedure, which utilizes different methods for estimation.

8 Conclusions

Fitting hierarchical logistic regression models to survey binary data is common in a number of disciplines. These models are useful in analyzing survey data in the presence of clustering or correlation, which otherwise would make standard approaches inappropriate due to the lack of independence amongst the outcomes. Although there are a number of powerful approaches for fitting these models, such as the GLIMMIX and NLMIXED procedures in SAS, the computational complexity of the algorithms can often lead to failures in convergence.

Through the use of simulations, we obtained useful information for exploring the reasons for non-convergence, as well as steps to avoid these issues. In particular, when using the GLIMMIX procedure, researchers should be careful in selecting predictors to include in the model. The inclusion of too many predictors can lead to convergence issues, regardless of whether these predictors are fixed or random. When many predictors must be included due to research or knowledge constraints and if the GLIMMIX procedure failures to converge, other options can be explored to

fit similar models. Because it utilizes different approaches, the NLMIXED procedure is a viable option for obtaining convergence in the mixed model setting when the GLIMMIX procedure fails. Recent advances, such as the %HPGLIMMIX macro can also be utilized as a remedy.

While we concentrated and presented results applicable only to the convergence issue in the GLIMMIX procedure for two-level hierarchical logistic regression models, we believe that these approaches can be readily adapted and expanded to explore different or more complex problems. In general, Monte-Carlo simulation offers a fast, and inexpensive avenue for investigating problems such as convergence, as well as appropriate solutions.

Acknowledgements This work is funded in part by the National Institutes of Health Alzheimer's Consortium Fellowship Grant, Grant No. NHS0007. The content in this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Austin, P. C. (2010). Estimating multilevel logistic regression models when the number of clusters is low: A comparison of different statistical software procedures. *The International Journal of Biostatistics*, *6*(1), 1–20.

Austin, P. C., Manca, A., Zwarenstein, M., Juurlink, D. N., & Stanbrook, M. B. (2010). A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *Journal of Clinical Epidemiology*, *63*(2), 142–153.

Ene, M., Leighton, E. A., Blue, G. L., & Bell, B. A. (2015). Multilevel models for categorical data using SAS PROC GLIMMIX: The Basics. *SAS Global Forum 2015 Proceedings*.

Hartzel, J., Agresti, A., & Caffo, B. (2001). Multinomial logit random effects models. *Statistical Modelling*, *1*(2), 81–102.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An application of a mixed effects location scale model for analysis of Ecological Momentary Assessment (EMA) data. *Biometrics*, *64*(2), 627–634.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between- and within subject variance in Ecological Momentary Assessment (EMA) data using mixed-effects location scale models. *Statistics in Medicine*, *31*(27), 3328–3336.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah: Lawrence Erlbaum Associates Inc.

Irimata, K. M., & Wilson, J. R. (2017). Identifying Intraclass correlations necessitating hierarchical modeling. *Journal of Applied Statistics*, accepted.

Kiernan, K., Tao, J., & Gibbs, P. (2012). Tips and strategies for mixed modeling with SAS/STAT procedures. *SAS Global Forum 2012 Proceedings*.

Kuss, O. (2002). Global goodness-of-fit tests in logistic regression with sparse data. *Statistics in Medicine*, *21*, 3789–3801.

Kuss, O. (2002). How to use SAS for logistic regression with correlated data. In *SUGI 27 Proceedings* (pp. 261–27).

Lesaffre, E., & Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random effects model: An example. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *50*(3), 325–335.

Longford, N. T. (1993). *Random coefficient models*. Oxford: Clarendon Press.
Maas, C. J. M., & Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, *46*(3), 427–440.
McMahon, J. M., Pouget, E. R., & Tortu, S. (2006). A guide for multilevel modeling of dyadic data with binary outcomes using SAS PROC NLMIXED. *Computational Statistics & Data Analysis*, *50*(12), 3663–3680.
National Institute of Population Research and Training (NIPORT). (2013). *Bangladesh demographic and health survey 2011*. NIPORT, MItra and Associates, ICF International: Dhaka Bangladesh, Calverton MD.
Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)135*(3), 370–384.
Newsom, J. T. (2002). A multilevel structural equation model for dyadic data. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(3), 431–447.
Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2012). *User's guide to WLwiN, Version 2.26.* Centre for Multilevel Modelling, University of Bristol. Retrieved from http://www.bristol.ac.uk/cmm/software/mlwin/download/2-26/manual-web.pdf.
Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park: Sage Publications.
Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks: Sage Publications.
Rodriquez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A (Statistics in Society)158*(1), 73–89.
SAS Institute Inc. (2013). *Base SAS® 9.4 Procedure guide: Statistical procedures* (2nd ed.). Cary, NC: SAS Institute Inc.
Schabenberger, O. (2005). Introducing the GLIMMIX procedure for generalized linear mixed models. *SUGI 30 Proceedings*, 196–30.
Shahian, D. M., Normand, S. L., Torchiana, D. F., Lewis, S. M., Pastore, J. O., Kuntz, R. E., et al. (2001). Cardiac surgery report cards: Comprehensive review and statistical critique. *The Annals of Thoracic Surgery*, *72*(6), 2155–2168.
Smyth, G. K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society, Series B*, *51*, 47–60.
Snijders, T. A. B., & Bosker, R. J. (1998). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks: Sage Publications.
Three-level multilevel model in SPSS. (2016). UCLA: Statistical Consulting Group. http://www.ats.ucla.edu/stat/spss/code/three_level_model.htm.
Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, *61*(3), 439–447.
Xie, L., & Madden, L. V. (2014). %HPGLIMMIX: A high-performance SAS macro for GLMM Estimation. *Journal of Statistical Software*, *58*(8).
Wilson, J. R., & Lorenz, K. A. (2015). *Modeling Binary correlated responses using SAS, SPSS and R*. New York: Springer International Publishing.
Wolfinger, D. (1999). Fitting nonlinear mixed models with the new NLMIXED procedure. In *Sugi 24 Proceedings* (pp. 278–284).