Correction of Retransformation Bias in Nonlinear Predictions on Longitudinal Data with Generalized Linear Mixed Models

Liu X[1,2]*, Freed MC[1,2] and McCutchan PK[1]

[1]DOD Deployment Health Clinical Center, Defense Centers of Excellence, Walter Reed National Military Medical Center, Silver Spring, Maryland 20910, USA
[2]Department of Psychiatry, F. Edward Hebert School of Medicine,Uniformed Services University of the Health Sciences, Bethesda, Maryland 20814, USA

Abstract

Researchers often encounter discrete response data in longitudinal analysis. Generalized linear mixed models are generally applied to account for potential lack of independence inherent in longitudinally data. When parameter estimates are used to describe longitudinal processes, random effects, both between and within subjects, need to be retransformed in nonlinear predictions on the response data; otherwise, serious retransformation bias can arise to an unanticipated extent. This study attempts to go beyond existing work by developing a retransformation method deriving statistically robust longitudinal trajectory of nonlinear predictions. Variances of population-averaged nonlinear predictions are approximated by the delta method. The empirical illustration uses longitudinal data from the Asset and Health Dynamics among the Oldest Old study. Our analysis compares three sets of nonlinear predictions of death rate at six time points, from the retransformation method, the best linear unbiased predictor, and the fixed-effects approach, respectively. The results demonstrate that failure to retransform the random components in generalized linear mixed models results in severely biased nonlinear predictions, as well as much reduced standard error approximates.

Keywords: Generalized linear mixed models; Longitudinal data; Mixed-effects logit regression; Random effects; Retransformation bias

Introduction

In longitudinal data analysis, researchers frequently encounter discrete response data. When the distribution of the response variable is not normal or the variance/covariance matrices are not homogeneous longitudinally, the use of linear mixed models can lead to erroneous parameter estimates and unrealistic predictions of the response. There are different types of discrete longitudinal data, such as binary, ordinal, count, and multinomial. While there are particular model specifications and estimating procedures for each data type, the underlying expressions and statistical inferences in many of those models can be generalized by following the tradition of generalized linear models, referred to as generalized linear mixed models. Generalized linear mixed models incorporate subject-specific random effects, thereby addressing dependence among subject-specific observations. A variety of approaches have ben advanced to yield statistically efficient, robust, and consistent estimators on parameters of generalized linear mixed models [1-7].

In displaying analytic results of generalized linear mixed models, parameter estimates are often not directly interpretable due to transformation of distributional functions. With specification of the subject-specific random effects, covariates' regression coefficients in generalized linear mixed models do not necessarily describe changes in the mean response in the study population, and the actual effects must be evaluated by averaging over the distribution of specified random components [8]. The issue of interpretability in the analytic results of generalized linear mixed models thus calls for nonlinear predictions given the covariates' values, the estimated regression coefficients, and the average of random effects. When one converts a transformed linear function to predict the marginal mean at the untransformed scale, normality of the random components in the linear predictor must be retransformed to a non-normal distribution [9,10]; otherwise serious retransformation bias can arise. Even if true values of the parameters are known, the analytic results from generalized linear mixed models cannot be converted to unbiased nonlinear predictions without appropriately retransforming the random components.

In this article, we display an efficient, robust method for nonlinear predictions in generalized linear mixed models that corrects for retransformation bias. In Section 2, we briefly review general specifications of generalized linear mixed models. In Section 3, we present the classical best linear unbiased predictor for nonlinear predictions as an ancillary presentation of generalized linear mixed models. This is followed by a description of the retransformation method in nonlinear predictions. Section 5 presents approximation of the variance-covariance matrix for nonlinear predictions. An example is provided in Section 6 comparing results of nonlinear predictions from various methods. In the final section, we discuss merits and remaining issues in the methods described in this article.

Specifications and Inferences

Generalized linear mixed models are simply an extension of the classical generalized linear modeling from univariate data to clustered measurements. In the longitudinal setting, let i denote subject i in a random sample of N subjects and $j=1, ..., n_i$ be the number of repeated measurements nested within i. The response variable Y_{ij} can be regarded as a discrete realization of a random variable Y_{ij} with mean μ_{ij} and variance $\text{var}(y_{ij})$. The function $g(y_{ij}, x_{ij})$ is a nonlinear function linking Y_{ij} to covariate vector X_{ij} and a $1 \times q$ vector of subject-specific random effects $b_i = \{b_{i1}, ..., b_{iq}\}^\text{T}$. As the random components are unobservable, it is desirable to specify generalized linear mixed models based on expectation [1], written as

$$\text{E}\left(y_{ij}\right) = g^{-1}\left\{\text{E}\left(X_{ij}^\text{T}\boldsymbol{\beta} + Z_{ij}^\text{T}b_i\right)\right\} = \mu_{ij}, \tag{1}$$

*Corresponding author: Liu X, Department of Psychiatry, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814, USA, Tel: 301-295-7198; E-mail: xian.liu@usuhs.edu

Where $E(y_{ij})$ is the expected value of Y_{ij}, $\boldsymbol{\beta}$ is an $M \times 1$ vector of unknown regression parameters to be estimated including a time factor, and $\boldsymbol{Z}_{ij}$ is a design block matrix. The $1 \times q$ random effects vector $\boldsymbol{b}_i$ is distributed as $N(0,\boldsymbol{G})$, where $\boldsymbol{G}$ is a $q \times q$ covariance matrix. The matrix $\boldsymbol{Z}_{ij}$ can contain time or other covariates whose association with the response is assumed to vary across subjects. With the specification of $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$, the elements in vector $y_i = y_{i1}, \ldots y_{ni}$, are thought to be conditionally independent. The linear predictor, $g(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i) = E(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^T\boldsymbol{b}_i)$, is conventionally denoted by η_{ij}.

With the specification of a link function in generalized linear mixed models, random errors for nonlinear functions depend on the mean function, and accordingly, the variance of y_{ij} is written as

$$\operatorname{var}(y_{ij}) = \operatorname{var}(\mu_{ij}) + E[\varphi \upsilon(\mu_{ij})]$$
$$= \operatorname{var}[g^{-1}(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^T\boldsymbol{b}_i)] + E\{\varphi\upsilon[g^{-1}(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^T\boldsymbol{b}_i)]\}, \quad (2)$$

Where υ is a specific variance function, and ϕ represents a scale factor for over-dispersion. Given this flexible specification, y_{ij} can follow a probability distribution other than multivariate normality. Equation (2) includes two distinctive variance components, between-subjects and within-subject. Given the specification of the variance function, the within-subjects variance cannot be specified freely in non-normal longitudinal data.

Equation (1) does not specify a within-subject error term, implying zero uncertainty given $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$. From Equation (2), however, it seems desirable to express y_{ij} as a conditional function by including an error term for addressing uncertainty [9]:

$$y_{ij} = g^{-1}(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^T\boldsymbol{b}_i + \varepsilon_{ij}) = g^{-1}(\tilde{\eta}_{i,}) = \tilde{\mu}_{ij}, \quad (3)$$

Where $\tilde{\mu}_{ij}$ is the conditional mean after accounting for the within-subject random error ε_{ij}. Correspondingly, $\tilde{\eta}_{ij}$ is defined as

$$g(y_{ij}|\boldsymbol{b}_i) = \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^T\boldsymbol{b}_i + \varepsilon_{ij} = \eta_{ij} + \varepsilon_{ij} = \tilde{\eta}_{ij}. \quad (4)$$

Specification of the variance matrix for within-subjects random errors depends on a specific link function, and this random term may be conveniently assumed to have local normality with property $\varepsilon_{ij}|\boldsymbol{b}_i \sim N(0, \sigma_{\varepsilon j}^2)$. Due to retransformation of ε_{ij} in expressing the expected value of y_{ij}, very often $\tilde{\mu}_{ij} \neq \mu_{ij}$, albeit $\tilde{\eta}_{ij} = E(\eta_{ij})$. Let $\tilde{\boldsymbol{\alpha}}$ be a correlation parameter, and $\boldsymbol{R}(\tilde{a})$ be the $n_i \times n_i$ matrix of $\tilde{\boldsymbol{\alpha}}$ describing the correlation pattern within subject i. Then, if within-subjects uncertainty is considered, the within-subject variance-covariance structure can be written as

$$\operatorname{var}(\varepsilon_i|\boldsymbol{b}_i) = \varphi \boldsymbol{A}_i^{1/2}(\tilde{\boldsymbol{\mu}}_i) \boldsymbol{R}_i(\tilde{a}) \boldsymbol{A}_i^{1/2}(\tilde{\boldsymbol{\mu}}_i), \quad (5)$$

where $\tilde{\boldsymbol{\mu}}_i$ is the vector of means over n_i time points, $\boldsymbol{A}_i(\tilde{\boldsymbol{\mu}}_i)$ is an $n_i \times n_i$ diagonal within-subject variance matrix containing elements υ, evaluated at $\tilde{\boldsymbol{\mu}}_i$, and $\boldsymbol{R}_i$ is an unknown matrix to be estimated. For analytic convenience and simplicity, the matrices in Equation (5) are assumed to be common to all subjects. When the subject-specific random effects are specified, $\boldsymbol{R}_i(\tilde{\boldsymbol{\alpha}})$ is often simplified as $\boldsymbol{R}\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix.

The variance-covariance matrix of the linear predictor, denoted by $V_i(\tilde{\eta}_i)$, can be written as

$$V_i(\tilde{\eta}_i) = \varphi \boldsymbol{A}_i^{1/2}(\eta_i) \boldsymbol{R}_i(\tilde{a}) \boldsymbol{A}_i^{1/2}(\eta_i) + \boldsymbol{Z}_i^T\boldsymbol{G}\boldsymbol{Z}_i. \quad (6)$$

The parameters $\boldsymbol{\beta}$ and $\boldsymbol{G}$ can be estimated by the maximum likelihood or the restricted maximum likelihood, or other Bayes-type estimators; and the random effects are usually approximated as posterior modes [2,4,5-7]. The variance-covariance matrix of within-subjects uncertainty, if specified, can be approximated by using the Hessian of the log likelihood function or of the log restricted likelihood function. The analytic results from such procedures, however, are not sufficiently interpretable until they are converted into nonlinear predictions.

Best Linear Unbiased Predictor

There is a variety of approximation methods for nonlinear predictions on longitudinal data [1,7,11]. Among those approximation methods, a popular approach is the best linear unbiased predictor, which uses estimates of the fixed effects and the predicted values of the random effects. In this method, the linear predictor in generalized linear mixed models can be written as the mean

$$\hat{\eta}_i(\boldsymbol{y}_i|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{b}}_i) = \boldsymbol{X}_{ij}^T\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_{ij}^T\hat{\boldsymbol{b}}_i, \quad (7)$$

where $\boldsymbol{y}_i = \{y_{i1}, \ldots, y_{in_i}\}$, $\hat{\boldsymbol{\beta}}$ is the maximum likelihood or the restricted maximum likelihood estimate of $\boldsymbol{\beta}$, and $\hat{\boldsymbol{b}}_i$ is the prediction of $\boldsymbol{b}_i$. The marginal likelihood can be fitted by averaging over the distribution of the unobserved random effect $\boldsymbol{b}_i$, with the corresponding joint likelihoods over N subjects written as

$$L(\boldsymbol{\beta}, \boldsymbol{G}, \varphi) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i, \varphi) f(\boldsymbol{b}_i|\boldsymbol{G}) d\boldsymbol{b}_i. \quad (8)$$

The maximum likelihood or the restricted maximum likelihood estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{G}}$, and $\hat{}$ are the values of $\boldsymbol{\beta}$, $\boldsymbol{G}$, and ϕ that maximize the above likelihood function. The random effect predictions can be obtained as the conditional mean of $\boldsymbol{b}_i$ given, $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{G}}$, and $\hat{\varphi}$:

$$\hat{\boldsymbol{b}}_i = E(\boldsymbol{b}_i|\boldsymbol{y}_i; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{G}}, \hat{\varphi}).$$

Among various approximation techniques for $\hat{\boldsymbol{b}}_i$, quadrature methods are regarded to generate accurate approximations of the integral specified in the marginal likelihood function [4,5,8,12]. This popular procedure first uses the likelihood calculations given in McGilchrist [11], and then the empirical Bayes estimates of the random effects to approximate the Gaussian quadrature integral [13]. The approximated integral as a marginal likelihood is then optimized for the fixed effects, and the fixed effect estimates are applied to produce the final prediction. In this approach, as the random effect prediction is treated as the fixed effect, some inherent variability is overlooked thereby causing retransformation bias in nonlinear predictions. Therefore, the best linear unbiased predictor is a partial empirical Bayes method in nonlinear predictions. Furthermore, if there is evidence that uncertainty is not negligible, the estimator $\hat{y}_i = g^{-1}(\tilde{\eta}_i)$ does not result in an accurate nonlinear predictor for y_i because $E[g^{-1}(\hat{\boldsymbol{\varepsilon}}_i)] \neq g^{-1}[E(\hat{\boldsymbol{\varepsilon}}_i)]$.

The marginal mean for a population subgroup can be predicted from subject-specific predictions by creating a scoring dataset that represents an actual or a hypothetical population group taking selected values of covariates. By retaining the predicted random effect for each subject, the mean and the standard deviation of subject-specific predictions in the scoring dataset approximate the population-averaged prediction and its standard error by use of the best linear unbiased predictor.

Retransformation Method

One limitation in the best linear unbiased predictor is that $\hat{\boldsymbol{b}}_i$ is

treated as the fixed effect in nonlinear predictions. Consequently, a portion of variability is ignored in the retransformation process. For a population group, the entire set of random components needs to be retransformed for correctly approximating the population-averaged mean. We refer to this retransforming of random components in nonlinear predictions as the retransformation method. In this section, we present two such methods, one without and one with considering within-subjects uncertainty, respectively.

Retransformation method without considering intra-subjects variability

We start the description with the case of a log link. Let subject i be a typical case in a population of interest characterized by covariate vector X_0. The predicted value for the typical subject at time j can then be considered the population-averaged prediction for the population taking covariate values X_0. Given Equation (1), the marginal mean of y_{ij} given X_{0j} and b_i is

$$
\begin{aligned}
\mathrm{E}\left(\hat{y}_{ij} \mid X_{0j}, b_i\right) &= \mathrm{E}\left[\exp\left(X_{0j}^{\mathrm{T}}\hat{\beta} + Z_{0j}^{\mathrm{T}}\hat{b}_i\right)\right] \\
&= \exp\left(X_{0j}^{\mathrm{T}}\hat{\beta}\right)\mathrm{E}\left[\exp\left(Z_{0j}^{\mathrm{T}}\hat{b}_i\right)\right] \\
&= \exp\left(X_{0j}^{\mathrm{T}}\hat{\beta}\right)\hat{\Phi}_{0j},
\end{aligned}
\tag{9}
$$

where $\hat{\Phi}_{0j} = \mathrm{E}\left[\exp\left(Z_{0j}^{\mathrm{T}}\hat{b}_i\right)\right]$ is defined as the moment generating function of b_i evaluated at Z_{0j}. Given $b_i \sim \mathrm{N}(0, G)$, Φ_{ij} is the expectation of a lognormal distribution, mathematically defined as $\exp\left\{\left(Z_{ij}G Z_{ij}^{\mathrm{rT}}\right)/2\right\}$. Therefore, the prediction at time j for the population with covariate X_{0j} is given by

$$
\mathrm{E}\left(\hat{y}_{ij} \mid X_{0j}, \hat{b}_i\right) = \exp\left(X_{0j}^{\mathrm{T}}\hat{\beta}\right)\exp\left(\frac{Z_{0j}\hat{G} Z_{0j}^{\mathrm{T}}}{2}\right).
\tag{10}
$$

Equation (10) indicates that $\exp\left(X_{0j}^{\mathrm{T}}\hat{\beta}\right)\exp\left\{\left(Z_{ij}\hat{G} Z_{ij}^{\mathrm{T}}\right)/2\right\} > \exp\left(X_{0j}^{\mathrm{T}}\hat{\beta}\right)$ unless all elements in $\hat{G}$ take value zero. Therefore, given the log link, the nonlinear response y_{ij} will be under-predicted if retransformation of between-subjects random effects is completely or partially neglected, with the magnitude of such retransformation bias depending on the size of $\hat{\Phi}_{0j}$.

Next, let g(·) represent the popular logit link. Given the logit function, neglect of retransformation of random effects can also lead to tremendous retransformation bias on the binary data. Let y_{ij} denote a dichotomous response variable taking value 0 or 1 for subject i at time j. Then the predicted probability that $y_{ij} = 1$ for person i at time j given X_{0j} and b_i can be written as

$$
\begin{aligned}
\mathrm{E}\left(\hat{p}_{ij} \mid X_{0j}, \hat{b}_i\right) &= \mathrm{E}\left[\mathrm{pr}\left(y_{ij} = 1 \mid X_{0j}, \hat{b}_i\right)\right] = \mathrm{E}\left\{\frac{1}{1 + \exp\left[-\left(X_{0j}^{\mathrm{T}}\hat{\beta} + Z_{0j}^{\mathrm{T}}\hat{b}_i\right)\right]}\right\} \\
&= \left[1 + \exp\left(-X_{0j}^{\mathrm{T}}\hat{\beta}\right)\mathrm{E}\left\{\exp\left(-Z_{0j}^{\mathrm{T}}\hat{b}_i\right)\right\}\right]^{-1} \\
&= \left\{1 + \exp\left(-X_{0j}^{\mathrm{T}}\hat{\beta}\right)\left(\frac{1}{\hat{\Phi}_{0j}}\right)\right\}^{-1},
\end{aligned}
\tag{11}
$$

where; in the construct of the lognormal distribution,

$$
\hat{\Phi}_{0j} = \exp\left(\frac{Z_{0j}\hat{G} Z_{0j}^{\mathrm{T}}}{2}\right).
$$

Let the fixed-effects estimator $\mathrm{E}\left(\hat{p}_{ij} \mid X_{0j}\right) = \left[1 + \exp\left(-X_{0j}^{\mathrm{T}}\hat{\beta}\right)\right]^{-1}$ be $\hat{p}_{ij}^{(1)}$

and the mixed-effects estimator $\mathrm{E}\left(\hat{p}_{ij} \mid X_{0j}, \hat{b}_i\right)$ be $\hat{p}_{ij}^{(2)}$. Then, when $\hat{\Phi}_{0j} > 1$, which is generally the case in longitudinal data analysis, we have $\hat{p}_{ij}^{(2)} > \hat{p}_{ij}^{(1)}$. Therefore, in nonlinear predictions with mixed-effects logit models, ignoring retransformation of between-subjects random effects can lead to strong downward bias in the predicted probability. For other discrete data types, such as the multinomial, the retransformation bias in nonlinear predictions can be equally impactful [9].

Retransformation method with intra-subjects variability

The equations specified in the above section do not specify a random term accounting for within-subjects uncertainty. In the application of generalized linear mixed models, between-subjects random effects are often perceived to reflect individual differences in unspecified characteristics thereby addressing within-subjects variability simultaneously [14]. In certain situations, however, this assumption does not reflect the true experiences generated by the stochastic longitudinal process.

If within-subjects variability is taken into account, the expectation of nonlinear response y for subject i at time j can be written as

$$
\mathrm{E}\left(y_{ij}\right) = \mathrm{E}\left\{g^{-1}\left[X_{ij}^{\mathrm{T}}\hat{\beta} + Z_{ij}^{\mathrm{T}}\hat{b}_i + \Delta\left(y_{ij} - \mu_{ij}\right)\right]\right\},
\tag{12}
$$

where $\Delta\left(y_{ij} - \mu_{ij}\right)$ can be understood as a second-order smearing estimate evaluated at $\left(\hat{\beta}, \hat{b}_i\right)$. Empirically, the within-subjects random term can be approximated from the partial derivative of the log likelihood function with respect to β in estimating a generalized linear mixed model, given by

$$
\hat{\Delta}\left(y_{ij} - \mu_{ij}\right) \approx \frac{\partial \log L\left(y_{ij} \mid \beta, G, \varphi\right)}{\partial\beta}\bigg|_{\hat{\beta}, \hat{b}}.
$$

Practically, the term $\hat{\Delta}\left(y_{ij} - \mu_{ij}\right)$ is the local approximation of the within-subjects random error with local normality. Such an approximate is model-based; different from linear mixed models, its specification depends on the marginal mean. In nonlinear predictions, the specification of this local approximation step can be ignored only when there is strong evidence that between-subjects variability completely captures within-subjects uncertainty.

Some researchers recommend the application of the latent variable approach to estimate within-subjects random errors in generalized linear mixed models [15]. This standardized approach specifies a constant variance of within-subjects random errors, regardless of the response type and the number of covariates utilized in a particular longitudinal study. It is argued that, when the between-subjects random effects are specified, not that much variability remains in a binary or a multinomial response [15]. Furthermore, this approach does not specify a covariance structure when the data type is multinomial thereby overlooking the multivariate nature of the nonlinear data structure.

If within-subjects random errors are considered non-ignorable in specifying a generalized linear mixed model, the marginal mean of the nonlinear response y_{ij}, taking the logit link, is

$$
\begin{aligned}
\mathrm{E}\left(\hat{p}_{ij} \mid X_{0j}, \hat{b}_i\right) &= \mathrm{E}\left[\frac{1}{1 + \exp\left\{-\left(X_{0j}^{\mathrm{T}}\hat{\beta} + Z_{0j}^{\mathrm{T}}\hat{b}_i + \hat{\Delta}_{ij}\right)\right\}}\right] \\
&= \left[1 + \exp\left(-X_{0j}^{\mathrm{T}}\hat{\beta}\right)\mathrm{E}\left\{\exp\left(-Z_{0j}^{\mathrm{T}}\hat{b}_i\right)\right\}\mathrm{E}\left\{\exp\left(-\hat{\Delta}_{ij}\right)\right\}\right]^{-1}
\end{aligned}
$$

$$= \left[1 + \exp\left(-\boldsymbol{X}_{0j}^{\mathrm{T}} \hat{\boldsymbol{\beta}}\right) \left\{ \exp\left(\frac{\boldsymbol{Z}_{0j} \hat{\boldsymbol{G}} \boldsymbol{Z}_{0j}^{\mathrm{T}} + \hat{\sigma}_{\varepsilon j}^2}{2} \right) \right\}^{-1} \right]^{-1}, \qquad (13)$$

where $\hat{\varDelta}_{ij} = \hat{\varDelta}\left(y_{ij} - \mu_{ij}\right)$ and $\hat{\sigma}_{\varepsilon j}^2$ is the variance of $\hat{\varDelta}_{ij}$ that can be obtained from the inverse of the observed information matrix for an underlying marginal model. Let the above estimator be $\hat{p}_{ij}^{(3)}$. Then, we have $\hat{p}_{ij}^{(3)} > \hat{p}_{ij}^{(2)} > \hat{p}_{ij}^{(1)}$ unless $\hat{\sigma}_{\varepsilon j}^2 = 0$. Therefore, if within-subject uncertainty cannot be ignored, overlooking retransformation of the within-subjects random component results in underestimation of the predicted probability.

Variance Matrix for Nonlinear Predictions

We propose to use the delta method to approximate the standard errors of nonlinear predictions, as following the tradition in nonlinear predictions [12-14,16,17]. Suppose that the discrete response is binomial. Let $\hat{\mathcal{L}}_{ij}$ be a random variable of the linear predictor for subject i at time point j from Equation (4) with mean $\tilde{\eta}_{ij}$ and variance $\mathrm{var}\left(\hat{\mathcal{L}}_{ij}\right)$, and $\hat{p}_{ij} = g^{-1}\left(\hat{\mathcal{L}}_{ij}\right)$ is a transform of $\hat{\mathcal{L}}_{ij}$, as defined by Equation (3). As conventionally defined, g is the link function and g^{-1} is its inverse function. For large samples, the first-order Taylor series expansion of $g^{-1}\left(\hat{\mathcal{L}}_{ij}\right)$ yields approximation of mean

$$E\left\{ g^{-1}\left(\hat{\mathcal{L}}_{ij}\right) \right\} \approx g^{-1}\left(\tilde{\eta}_{ij}\right), \qquad (14)$$

and; of variance $\mathrm{var}\left(\hat{p}_{ij}\right)$

$$\mathrm{var}\left\{ g^{-1}\left(\hat{\mathcal{L}}_{ij}\right) \right\} \approx \left\{ \frac{\partial\, g^{-1}\left(\hat{\mathcal{L}}_{ij}\right)}{\partial\, \hat{\mathcal{L}}_{ij}} \bigg| \hat{\mathcal{L}}_{ij} = \tilde{\eta}_{ij} \right\}^2 \mathrm{var}\left(\hat{\mathcal{L}}_{ij}\right). \qquad (15)$$

In Equation (15), $\mathrm{var}\left\{ g^{-1}\left(\hat{\mathcal{L}}_{ij}\right) \right\}$ is the approximation for the variance of the predicted probability $\hat{p}_{ij}$ for large samples. The square-root of this variance approximation yields the standard error of the nonlinear prediction given $\hat{}$, $\hat{\boldsymbol{G}}$, and $\hat{\varphi}$.

If the discrete response is multinomial, the two scalars $\hat{\mathcal{L}}_{ij}$ and $g^{-1}\left(\hat{\mathcal{L}}_{ij}\right)$ need to be replaced by two random vectors, $\hat{\boldsymbol{\mathcal{L}}} = \hat{\mathcal{L}}_{ij1}, \dots, \hat{\mathcal{L}}_{ijK}$ and $\hat{\boldsymbol{p}} = g^{-1}\left(\hat{\mathcal{L}}_{ij1}, \dots, \hat{\mathcal{L}}_{ijK}\right)$, respectively, where K denotes the number of non-reference response levels [9]. Correspondingly, the variance-covariance matrix $V\left(\hat{\boldsymbol{p}}\right)$ is approximated by

$$V\left[g^{-1}\left(\hat{\boldsymbol{\mathcal{L}}}\right) \right] \approx \left[\frac{\partial\, g^{-1}\left(\hat{\boldsymbol{\mathcal{L}}}\right)}{\partial\, \hat{\boldsymbol{\mathcal{L}}}} \bigg| \hat{\boldsymbol{\mathcal{L}}} = \tilde{\boldsymbol{\eta}} \right]^{\mathrm{T}} \boldsymbol{\Sigma}\left(\hat{\boldsymbol{\mathcal{L}}}\right) \left[\frac{\partial g^{-1}\left(\hat{\boldsymbol{\mathcal{L}}}\right)}{\partial\, \hat{\boldsymbol{\mathcal{L}}}} \bigg| \hat{\boldsymbol{\mathcal{L}}} = \tilde{\boldsymbol{\eta}} \right], \qquad (16)$$

where;

$$\frac{\partial\, g^{-1}\left(\hat{\boldsymbol{\mathcal{L}}}\right)}{\partial\, \hat{\boldsymbol{\mathcal{L}}}} = \left[\frac{\partial g_1^{-1}\left(\hat{\boldsymbol{\mathcal{L}}}\right)}{\partial \hat{\boldsymbol{\mathcal{L}}}}, \frac{\partial g_2^{-1}\left(\hat{\boldsymbol{\mathcal{L}}}\right)}{\partial \hat{\boldsymbol{\mathcal{L}}}}, \dots \right],$$

and

$$\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\mathcal{L}}}\right) = \begin{pmatrix} \mathrm{var}\left(\hat{\mathcal{L}}_1\right) & \mathrm{cov}\left(\hat{\mathcal{L}}_1, \hat{\mathcal{L}}_2\right) & \cdots & \mathrm{cov}\left(\hat{\mathcal{L}}_1, \hat{\mathcal{L}}_K\right) \\ & \mathrm{var}\left(\hat{\mathcal{L}}_2\right) & \cdots & \mathrm{cov}\left(\hat{\mathcal{L}}_2, \hat{\mathcal{L}}_K\right) \\ & & \ddots & \\ & & & \mathrm{var}\left(\hat{\mathcal{L}}_K\right) \end{pmatrix}.$$

In generalized linear mixed models, $\mathrm{var}\left(\hat{\mathcal{L}}_{ij}\right)$ may consist of two components, between-subjects and within-subjects variances, as indicated earlier. Whereas the between-subjects variance is specified in a generalized linear mixed model and thereby estimable, the within-subjects random component can be approximated from the variance of the intercept after the covariates are rescaled to be centered at selected values [9]. The rationale is that if covariates are centered at some specified values, the intercept represents the transformed margin, and therefore, the variance of the estimated intercept plus the corresponding variance for the between-subjects random effects can be considered the approximation of the variance for the marginal mean.

Illustration

Data used for the illustration come from the Survey of Asset and Health Dynamics among the Oldest Old, a nationally representative investigation of older Americans. This survey, conducted by the Institute for Social Research, University of Michigan, is funded by National Institute on Aging as a supplement to the Health and Retirement Study. To date, the survey consists of nine waves of data collection. The Wave I survey was conducted between October 1993 and April 1994, identifying 9,473 households and 11,965 individuals in the target area range. The survey obtains detailed information on a number of domains, including demographics, health status, health care use, disability, and health and life insurance. Survival information throughout the follow-up waves is obtained through a link to the data of National Death Index. Because the first two waves (1993 and 1995) were based on a slightly different questionnaire from those of the succeeding waves, we use data from the six waves starting with the 1998 wave (1998, 2000, 2002, 2004, 2006, and 2008). For details about the study design, see the Health and Retirement Study website (hrsonline. isr.umich.edu).

In this illustration, the outcome variable is survival status, with 1=death and 0=else. Operationally, we analyze the probability of death between two successive waves, defined as $\mathrm{pr}(Y_{ij}=1)$ where j indicates a time interval between time point $j-1$ and j. Given the focus on nonlinear predictions on longitudinal trajectory of death rate, the main explanatory variable is time, measured as the number of years elapsed since the 1998 survey ($t = 0, 2, 4, 6, 8, 10$). Among the control variables considered in this illustration, gender is a dichotomous variable with 1=women and 0=men, and age and education are both continuous variables. In estimating the model parameters, the control variables are rescaled to be centered about sample means. The interaction between time and each of the controls was considered for capturing possible convergence of the covariate's effect, but its inclusion did not contribute significantly to the model fit and thus was removed.

Given the binary outcome data, we apply the mixed-effects binary logit model. For illustrative simplicity without loss of generality, we use the random intercept logit model, assuming the effects of covariates on the logit to be fixed over time. The SAS PROC NLMIXED procedure (SAS Institute Inc., Cary, NC) is applied to compute the fixed and the random effects given its tremendous flexibility in estimating and predicting parameters in generalized linear mixed models [16]. We use adaptive Gaussian quadrature to approximate the integral of the likelihood over the random intercept with its advantage over other approximation methods in deriving robust random effect estimates and the model fit statistic [4,12]. With the specification of between-subjects random intercepts, time is treated as a continuous variable. A combination of time and time × time, the so-called quadratic polynomial time function, was found to best describe the longitudinal

trajectory of death rate. Given high correlation between the two time components, the time variable t was rescaled as a centered covariate to reduce numeric instability and collinearity.

To compare statistical efficiency and robustness of different methods for nonlinear predictions in longitudinal analysis, we first create a "full" logit model consisting of all covariates and two random components in the estimation process, assuming the population-averaged predictions and the corresponding variances/covariances derived from this model to be exact. Then we purposefully exclude the variable "education" from the logit model. As education has a statistically significant effect on death rate, there is definitely additional clustering in the outcome data after its removal, and therefore such a reduced model is actually a misspecified model. We have three operational objectives in this illustration. First, we examine whether the retransformation method can capture the random effect after an important predictor is removed, relative to the results from the full model. Next, we assess whether the retransformation method yields much smaller retransformation bias than the best linear unbiased predictor in nonlinear predictions. Third, we demonstrate how the fixed-effects approach, though tending to generate unbiased estimates of regression coefficients [18], results in tremendous retransformation bias in nonlinear predictions. The best linear unbiased predictor and the retransformation method are actually based on the same model eliminating education from the estimating process, and therefore, we develop three logit models in the illustration: the full mixed-effects logit model, the reduced mixed-effects logit model removing education from regression, and the reduced fixed-effects logit model.

While the best linear unbiased predictor approximates death rate for each observation, we obtain the population-averaged predictions from this method by creating a scoring dataset that specifies the outcome variable as missing and the independent variables as zero, with covariates rescaled to be centered at selected values corresponding to a target population group. In the scoring data, the random effect per individual should be present, and therefore, the mean of the predictions with y=missing and X's=0 in the scoring dataset generates the predicted death rate for the population or a typical subject taking selected values of covariates.

Table 1 displays the analytic results of three models. All regression coefficients, full or reduced, random- or fixed-effects, are statistically significant at α=0.05. In each model, the regression coefficient of time on the logit is positive, while that of time × time is negative, which, combined, suggest a decelerating time trend in old-age mortality. The between-subjects random effect on the logit is 0.381 for the full model and 0.380 for the reduced, both statistically significant. The intercept, the regression coefficients, and the standard errors in the reduced fixed-effects model are close to those from the two random-effects logit models. Such a similarity indicates that in this analysis, the large-sample behavior follows, and consequently, the asymptotic process $n^{1/2}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{0j}\right)$, where $\boldsymbol{\beta}_0$ is the true parameter vector, tends to converge to a normal vector with mean **0** and the covariance matrix as approximated by the inverse of the observed information matrix. The fixed-effects logit model, however, does not have capability to yield a robust and consistent estimator for nonlinear predictions in the longitudinal setting, as will be shown in Table 1.

The predicted death rate at each time point and its variance can be estimated by applying Equations (14) and (15), respectively. In the construct of binary response data, the partial derivative of death rate with respect to the logit function is

Explanatory variable and effect type	Regression coefficient	Standard error
Full random-effects logit model		
Fixed Effects:		
Intercept	-1.486***	0.033
Time (centered)	0.292***	0.016
Time × time (centered)	-0.047***	0.002
Gender (centered)	-0.476***	0.052
Age (centered)	0.132***	0.006
Education (centered)	-0.032***	0.006
Random Effects:		
Intercept	0.381***	0.017
Reduced random-effects logit model		
Fixed Effects:		
Intercept	-1.472***	0.031
Time (centered)	0.271***	0.018
Time × time (centered)	-0.047***	0.002
Gender (centered)	-0.446***	0.051
Age (centered)	0.129***	0.006
Random Effects:		
Intercept	0.380***	0.017
Reduced fixed-effects logit model		
Fixed Effects:		
Intercept	-1.448***	0.027
Time (centered)	0.229***	0.007
Time × time (centered)	-0.045***	0.002
Gender (centered)	-0.390***	0.040
Age (centered)	0.116***	0.003

***p<0.01.

Note: Randomness of the intercept is parameterized by the variance of the random effects.

Table 1: Analytic results of three binary logistic regression models on the inter-wave death rate in older Americans (N=7,480).

$$\left\{\frac{\partial\, g^{-1}\left(\hat{\mathcal{L}}_{ij}\right)}{\partial\,\hat{\mathcal{L}}_{ij}}\bigg|\hat{\mathcal{L}}_{ij}=\tilde{\eta}_{ij}\right\} = \frac{\exp\left(\hat{\mathcal{L}}_{ij}\right)}{\left\{1+\exp\left(\hat{\mathcal{L}}_{ij}\right)\right\}^2}.$$

The variance of the predicted population-averaged death rate $\hat{p}_{ij}$ can be approximated by

$$\mathrm{var}\left(\hat{p}_{ij}\right) \approx \left\{\frac{\partial\, g^{-1}\left(\hat{\mathcal{L}}_{ij}\right)}{\partial\,\hat{\mathcal{L}}_{ij}}\bigg|\hat{\mathcal{L}}_{ij}=\tilde{\eta}_{ij}\right\}^2 \mathrm{var}\left(\hat{\mathcal{L}}_{ij}\right) = \left[\frac{\exp\left(\hat{\mathcal{L}}_{ij}\right)}{\left\{1+\exp\left(\hat{\mathcal{L}}_{ij}\right)\right\}^2}\right]^2 \mathrm{var}\left(\hat{\mathcal{L}}_{ij}\right),$$

where $\mathrm{var}\left(\hat{\mathcal{L}}_{ij}\right)$ consists of two components if within-subjects random errors are considered.

Table 2 presents four sets of the predicted death rates and the standard errors at six time points, computed from the full model, the retransformation method, the best linear unbiased predictor, and the fixed-effects reduced model, respectively.

In Table 2, death rate is shown to increase exponentially over time at the early and the middle stage, and then the increase slows near the end of the observation period, reflecting a "selection of the fittest" process. The retransformation method, transferring both the between-subjects and the within-subject random components in this analysis, generates the closest predictions to those from the full model, indicating its high efficiency and coverage in handling retransformation bias in nonlinear predictions. In the first two panels, the predicted

Type of Statistics	Inter-wave intervals				
	1998-2000	2000-2002	2002-2004	2004-2006	2006-2008
	Predicted death rates generated from the full mixed-effects model				
Death rate	0.075	0.169	0.259	0.294	0.254
Standard error	0.043	0.087	0.119	0.128	0.118
	Predicted death rates generated from the retransformation method				
Death rate	0.075	0.168	0.258	0.291	0.252
Standard error	0.043	0.086	0.118	0.128	0.117
	Predicted death rates generated from the best linear unbiased predictor				
Death rate	0.064	0.146	0.227	0.257	0.221
Standard error	0.012	0.024	0.034	0.037	0.034
	Predicted death rates generated from the reduced fixed-effects model				
Death rate	0.073	0.152	0.220	0.238	0.194
Standard error	0.002	0.003	0.005	0.005	0.008

Note: Except the first model, all other methods in Table 2 are based on "reduced" (misspecified) models.

Table 2: Predicted inter-waves death rates and standard errors at five longitudinal intervals from 1998 to 2008: AHEAD longitudinal survey (N=7,480).

death rates in the first two observation intervals, 1998-2000 and 2000-2002, are not statistically significant at $\alpha=0.05$, with high standard errors. The best linear unbiased predictor provides close predictions at the early times; in the last three time periods, however, the predicted death rate deviates markedly from those from the first two methods. Furthermore, this method results in severely underestimated standard errors of the predictions. The fixed-effects logit model generates the least valid predictions with massive deviations from those of the mixed-effects models. More significantly, the fixed-effects approach severely underestimates standard errors, much more so than the best linear unbiased predictor. With severe underestimation of standard errors, all predicted death rates from the best linear unbiased predictor and the fixed-effects model are very strongly statistically significant, thereby yielding erroneous test results.

Figure 1 plots three longitudinal trajectories of death rate and the corresponding 95% confidence limits, approximated from the retransformation method, the best linear unbiased predictor, and the fixed-effects approach, respectively. The solid line and the shaded area in each panel are the trajectory and the 95% confidence region from the full model, used as a standard for comparison. Panel A demonstrates how well the retransformation method predicts old-age mortality and the confidence limits, after one theoretically important, statistically significant predictor is removed. In Panel B, the best linear unbiased predictor estimates death rate fairly nicely in the early stage, but it then falls off systematically from the solid line, with the 95% confidence limits dramatically contracted. In Panel C, the predicted time trend of death rate is completely amiss, with the two 95% confidence limits very narrowly scattered around the predicted curve.

Discussion

As regression coefficients in generalized linear mixed models are often not interpretable, we have seen applications using the fixed-effect estimates for nonlinear predictions. Such an application can lead to considerable retransformation bias due to neglect of retransforming random components. If random disturbances in a generalized linear mixed model truly follow normality, this distribution needs to convert to a non-normal function to correctly predict the nonlinear outcome. Consequently, the expectation of random components is not zero in nonlinear predictions unless the identity link function is specified or both between- and within-subjects random disturbances are zero for all subjects. Without appropriately retransforming the random effects,
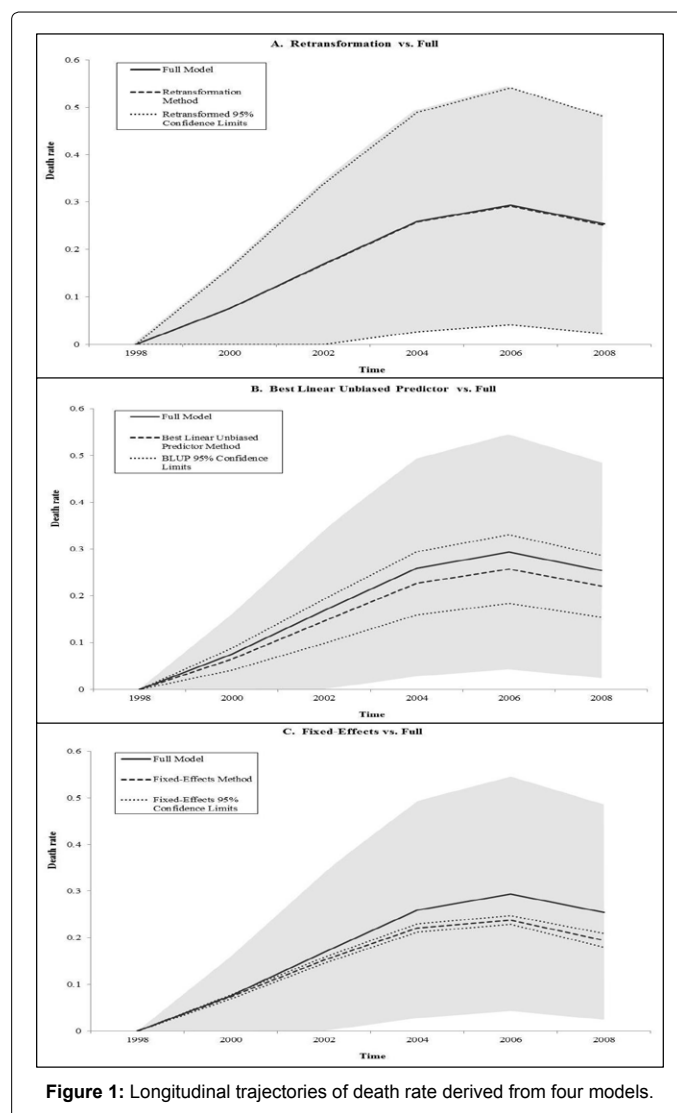

Figure 1: Longitudinal trajectories of death rate derived from four models.

the variance of nonlinear predictions will be underestimated, thereby affecting the quality of the significance test. The best linear unbiased

prediction, widely applied in longitudinal analysis, only accounts for a portion of between-subjects variability and overlooks the within-subject random component that may exist inherently even with the specification of the between-subjects random effects.

In this article, we compare the predicted death rate at six time points from three predicting methods: the retransformation method, the best linear unbiased predictor, and the fixed-effects approach, as relative to a pre-specified full mixed-effects logit model. In particular, we examine how each of these methods behaves in nonlinear predictions after an important predictor variable is removed. The results of our illustration display that failure to retransform random components in generalized linear mixed models can result in severe bias in nonlinear predictions and sizable underestimation of standard errors, even when the estimated regression coefficients are unbiased. Such retransformation bias exerts substantial influences on the quality of significance test on nonlinear predictions. The fixed effects are population-averaged, and therefore, the subject-specific variability is disregarded; once random effects are considered in nonlinear predictions, the inherent variability is much increased, thereby lowering the value of the chi-square statistic. The best linear unbiased predictor reduces some retransformation bias but its effect is shown to be very limited. Relative to the fixed-effects approach and the BLUP estimator that are associated with tremendous retransformation bias, our retransformation method is shown to increase the efficiency and coverage in nonlinear predictions.

Acknowledgement

References

1. Breslow NR, Clayton DG (1993) Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88: 9-25.

2. Davidian M, Giltinan DM (1995) Nonlinear models for repeated measurement data. Chapman and Hall/CRC, New York.

3. Little R JA, Rubin DB (2002) Statistical analysis with missing data. (2ndedn.) Wiley, New York.

4. McCulloch CE, Searle S, Neuhaus JM (2008) Generalized, linear, and mixed models. Wiley, Hoboken, New Jersey.

5. Pinheiro JC, Bates DM (1995) Approximation to the log-likelihood function in the nonlinear mixed effects model. Journal of Computational and Graphical Statistics 4: 12-35.

6. Stiratelli R, Laird N, Ware JH (1984) Random-effects models for series observations with binary response. Biometrics 40: 961-971.

7. Zeger SL, Karim MR (1991) Generalized linear models with random effects: A Gibbs sampling approach. Journal of the American Statistical Association 86: 79-86.

8. Fitzmaurice GM, Laird NM, Ware JH (2004) Applied longitudinal analysis. Wiley, Hoboken, New Jersey.

9. Liu X, Engel CC (2012) Predicting longitudinal trajectories of health probabilities with random-effects multinomial logit regression. Statistics in Medicine 31: 4087-4101.

10. Zeger SL, Liang KY, Albert PS (1988) Models for longitudinal data: A generalized estimating equation approach. Biometrics 44: 1049-1060.

11. McGilchrist CA (1994) Estimation in generalized mixed models. Journal of the Royal Statistical Society Series B 56: 61- 69.

12. Molenberghs G, Verbeke G (2010) Models for discrete longitudinal data. Springer, New York.

13. SAS (2009) SAS/STAT 12.1: User's guide. SAS Institute, Cary, North Carolina.

14. Diggle PJ, Heagerty PJ, Liang KY, Zeger SL (2002) Analysis of longitudinal data (2ndedn.) Clarendon Press, Oxford.

15. Hedeker D, Gibbons RD (2006) Longitudinal data analysis. Wiley, Hoboken, New Jercey.

16. Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Schabenberger O (2006) SAS for mixed models (2ndedn.) SAS Institute, Gary, North Carolina.

17. Stuart A, Ord JK (1994) Kendall's advanced theory of statistics (6thedn.) Edward Arnold, New York.

18. Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73: 13-22.