Biometrics 74, 354–361
March 2018 DOI: 10.1111/biom.12707

# A Note on Marginalization of Regression Parameters from Mixed Models of Binary Outcomes

**Donald Hedeker,[1,*] Stephen H. C. du Toit,[2] Hakan Demirtas,[3] and Robert D. Gibbons[1]**

[1]Department of Public Health Sciences, University of Chicago, 5841 S. Maryland Avenue, Room W254, MC2000, Chicago, Illinois 60637, U.S.A.
[2]Scientific Software International, Inc., Skokie, Illinois 60076, U.S.A.
[3]School of Public Health, University of Illinois at Chicago, Chicago, Illinois 60612, U.S.A.
**email:* hedeker@uchicago.edu

Summary. This article discusses marginalization of the regression parameters in mixed models for correlated binary outcomes. As is well known, the regression parameters in such models have the "subject-specific" (SS) or conditional interpretation, in contrast to the "population-averaged" (PA) or marginal estimates that represent the unconditional covariate effects. We describe an approach using numerical quadrature to obtain PA estimates from their SS counterparts in models with multiple random effects. Standard errors for the PA estimates are derived using the delta method. We illustrate our proposed method using data from a smoking cessation study in which a binary outcome (smoking, Y/N) was measured longitudinally. We compare our estimates to those obtained using GEE and marginalized multilevel models, and present results from a simulation study.

Key words: Clustered data; Longitudinal data; Multilevel models; Population-averaged estimates; Subject-specific estimates.

## 1. Introduction

Mixed-effects models are a popular approach for analysis of longitudinal and/or clustered binary outcomes, with many texts either devoted to or with substantial treatment on the topic (Raudenbush and Bryk, 2002; Demidenko, 2004; Skrondal and Rabe-Hesketh, 2004; Molenberghs and Verbeke, 2005; Brown and Prescott, 2006; Hedeker and Gibbons, 2006; McCulloch et al., 2008; Fitzmaurice et al., 2011; Goldstein, 2011). Such models are often referred to as belonging to the class of models known as generalized linear mixed models (GLMMs). It is well known that the regression parameters in such models have the "subject-specific" (SS) or conditional interpretation (Neuhaus et al., 1991), representing effects of the covariates controlling for, or holding constant, the random effects. This is in contrast to the marginal or "population-averaged" (PA) estimates of, most notably, the GEE approach (Zeger and Liang, 1986) and the marginalized multilevel model (Heagerty and Zeger, 2000). The SS and PA regression parameters generally differ in meaning and value, and the SS regression parameters depend on how many random effects are included in the model (e.g., conditional on random intercepts only or conditional on random intercepts and trends). Thus, Zeger et al. (1988) recommend that SS models are useful when the focus is on making inference at the individual level, whereas PA models are most useful for population-based inference. Here, we describe a numerical approach to obtain PA estimates and standard errors from their SS counterparts for models with one or multiple random effects. We focus on longitudinal binary outcomes and present results primarily for the logit link function, but our approach applies to other GLMMs as well.

Analysts often decide a priori on conducting a PA or SS analysis depending on their target of inference. Thus, if one is interested in obtaining PA estimates, then typically a GEE or marginalized multilevel model is used. As is well known for longitudinal data, GEE estimates are valid under the restrictive missing data mechanism of missing completely at random (MCAR). Of course, weighted GEE (Robins et al., 1995) can be used to allow for missing at random (MAR), but this requires a joint modeling of the longitudinal outcome and the missing data, and further requires that the missing data model is correctly specified (Preisser et al., 2002; Demirtas, 2004). Alternatively, our approach uses the SS estimates from a full-likelihood mixed model analysis to obtain their PA counterparts, and therefore provides valid inference under MAR, though this does assume that the variance-covariance structure is correctly specified. Also, mixed models are more flexible in handling measurements made at irregular time intervals, since the covariance structure can depend on a continuous time variable. Marginalized multilevel models are a viable alternative to obtain PA estimates under MAR, and can allow for irregular time intervals. Available software for such marginalized multielevel models is somewhat limited, though Griswold et al. (2013) includes SAS and R scripts for estimating such models.

As an example, Gruder et al. (1993) report on the effects of a televised smoking cessation intervention in which nearly 500 smokers were randomized to one of three conditions and then followed longitudinally in terms of their binary smoking status (Y/N). The three conditions were no-contact control, discussion, and social support, and subjects were assessed by self report if they were smoking or not at post-intervention, 6-, 12-, and 24-months. The analysis in the paper consisted of a random intercept and trend probit model, which found that there were strong group effects at post-intervention that were not maintained over the two-year follow-up. The estimates presented in the paper were SS estimates, and therefore conditional on the random subject intercepts and trends. Alternatively, PA estimates would help to describe the magnitude of, for example, the post-intervention group differences in the population of subjects, rather than conditional on subject effects.

In terms of the organization of the article, Section 2 briefly describes mixed and marginal models for longitudinal binary data, Section 3 describes the proposed marginalization approach, Section 4 presents results of simulation studies, Section 5 presents a re-analysis of the Gruder et al. (1993) smoking cessation data, and Section 6 discusses some aspects of the proposed approach. Supplementary materials are mentioned in Section 7, including the smoking cessation dataset and a SAS syntax file that can be used to perform the proposed marginalization.

## 2. Regression Models for Longitudinal Binary Outcomes

Consider the mixed logistic regression model for a binary outcome $Y_{ij}$ from subject $i$ $(i = 1, \ldots, N)$ at timepoint $j$ $(j = 1, \ldots, n_i)$:

$$\log\left[\frac{P(Y_{ij}=1 \mid \boldsymbol{\upsilon}_i)}{1-P(Y_{ij}=1 \mid \boldsymbol{\upsilon}_i)}\right] = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\upsilon}_i \ , \tag{1}$$

where the $r$ random effects $\boldsymbol{\upsilon}_i$ are normally distributed in the population with mean 0 and variance-covariance matrix $\boldsymbol{\Sigma}_\upsilon$. The total number of observations is denoted as $n = \sum^N n_i$. Using the Cholesky factorization of the random-effects variance-covariance matrix $\boldsymbol{\Sigma}_\upsilon = \boldsymbol{T}\boldsymbol{T}'$, the model can also be written as

$$\log\left[\frac{P(Y_{ij}=1 \mid \boldsymbol{\theta}_i)}{1-P(Y_{ij}=1 \mid \boldsymbol{\theta}_i)}\right] = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{T}\boldsymbol{\theta}_i \ , \tag{2}$$

where $\boldsymbol{\theta}_i$ are standardized random effects with mean 0 and variance-covariance matrix $\boldsymbol{I}$. The regression parameters $\boldsymbol{\beta}$ in this model have the "subject-specific" interpretation (Neuhaus et al., 1991), and so can be more precisely denoted as $\boldsymbol{\beta}^{ss}$ to reflect the fact that they represent effects of the covariates controlling for, or holding constant, the random subject effects. Thus, the meaning of $\boldsymbol{\beta}^{ss}$ depends on how many and which random effects are included in a given analysis.

In contrast to the mixed model, consider the marginal model

$$\log\left[\frac{P(Y_{ij}=1)}{1-P(Y_{ij}=1)}\right] = \boldsymbol{x}'_{ij}\boldsymbol{\beta}^{pa} \ , \tag{3}$$

in which there are no random effects. Instead, marginal models separate the mean and association structures. The model above is for the mean structure and would be combined with some specification of the association of the longitudinal data, for example, the specification of the working correlation in a GEE model. The regression parameters $\boldsymbol{\beta}^{pa}$ in this model have the marginal or "population-averaged" interpretation (Neuhaus et al., 1991), as they are not conditional on subject effects. As mentioned, these are preferable in settings in which one wants to make inferences about the population, for example in epidemiological studies. From this model, population-averaged probabilities, defined as $\pi^{pa}_{ij}$, can be expressed via the logistic response function as $\pi^{pa}_{ij} = \Psi(\lambda^{pa}) = 1/(1+\exp(-\lambda^{pa}))$, which translates the population-averaged logit $\lambda^{pa}$ into a probability $\pi^{pa}$. $\Psi$ is the CDF of the standard logistic distribution and is sometimes termed the "expit" function.

## 3. Marginalization of the Subject-Specific Regression Estimates

Suppose that one has estimated a mixed logistic regression model, with estimates $\hat{\boldsymbol{\beta}}^{ss}$ and $\hat{\boldsymbol{T}}$ (or $\hat{\sigma}_\upsilon$ for a random intercept model), and that one would also like to obtain the population-averaged regression coefficient estimates, denoted as $\hat{\boldsymbol{\beta}}^{pa}$. For a random-intercept probit model, there is a closed form solution, namely $\hat{\boldsymbol{\beta}}^{pa} = \hat{\boldsymbol{\beta}}^{ss}/\sqrt{1+\hat{\sigma}^2_\upsilon}$ (Heagerty and Zeger, 2000), while for a random-intercept logistic model, the following approximation can be used: $\hat{\boldsymbol{\beta}}^{pa} \approx \hat{\boldsymbol{\beta}}^{ss}/\sqrt{\frac{\hat{\sigma}^2_\upsilon+\pi^2/3}{\pi^2/3}}$, where $\pi^2/3$ is the variance of the standard logistic distribution (Agresti, 2002). Zeger et al. (1988) recommend multiplying this logistic variance by the factor 15/16, which results from using a cumulative Gaussian approximation to the logistic function. The square-root term on the right-hand side represents the ratio of the conditional to the marginal standard deviations, and can be viewed as the "marginalization" factor which transforms the subject-specific parameters into their population-averaged counterparts. In a random-intercepts model the marginalization factor is a scalar and equal across time. However, for models with multiple random effects this is generally not the case and so there is no simple relationship like the one above.

In the more general case of multiple random effects, following the results of Griswold and Zeger (2004) and others, population-averaged estimated probabilities for each observation can be expressed as

$$\hat{\pi}^{pa}_{ij} = \int_{\boldsymbol{\theta}} \Psi(\boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}}^{ss} + \boldsymbol{z}'_{ij}\hat{\boldsymbol{T}}\boldsymbol{\theta}_i)\, dF(\boldsymbol{\theta}) \ . \tag{4}$$

Gauss–Hermite quadrature can be used to approximate these probabilities as

$$\hat{\pi}^{pa}_{ij} \approx \sum_{q=1}^{Q^r} \Psi(\boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}}^{ss} + \boldsymbol{z}'_{ij}\hat{\boldsymbol{T}}\boldsymbol{B}_q) A(\boldsymbol{B}_q) \ . \tag{5}$$

The integration is approximated by a summation on a specified number of quadrature points $Q$ for each dimension of the integration. For the standard normal univariate density, optimal points and weights (which are denoted $B_q$ and $A(B_q)$, respectively) are given in Stroud and Sechrest (1966). For the multivariate density, the $r$-dimensional vector of quadrature points is denoted by $\boldsymbol{B}'_q = (B_{q1}, B_{q2}, \ldots, B_{qr})$, with its associated (scalar) weight given by the product of the corresponding univariate weights, $A(\boldsymbol{B}_q) = \prod_{h=1}^{r} A(B_{qh})$ .

Consider the marginal model in equation (3), and denote $\boldsymbol{\pi}^{pa}$ as the $n \times 1$ vector of probabilities that is obtained by stacking all of the $n$ observations, and $\boldsymbol{X}$ as the $n \times p$ covariate matrix that is obtained by stacking the covariate vectors of all $n$ observations (including a column of ones for the intercept). The marginal model can then be written as $\boldsymbol{\lambda}^{pa} = \boldsymbol{X}\boldsymbol{\beta}^{pa}$, where the $n \times 1$ vector of logits $\boldsymbol{\lambda}^{pa}$ is formed using elementwise division of $\boldsymbol{\pi}^{pa}$ by $1 - \boldsymbol{\pi}^{pa}$. Multiplying both sides of this equation by $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, yields the relationship between the population-averaged regression coefficients and the population-averaged logits:

$$\boldsymbol{\beta}^{pa} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\boldsymbol{\lambda}^{pa} \ . \tag{6}$$

Marginalized regression coefficients can be obtained using equation (6), in conjunction with the quadrature method in equation (5) which provides estimates of the marginal probabilities for each observation. Specifically, using estimates $\hat{\boldsymbol{\beta}}^{ss}$ and $\hat{\boldsymbol{T}}$, along with the known covariate and random effect matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$, in equation (5) yields $\hat{\pi}_{ij}^{pa}$ from which the $n \times 1$ vector $\hat{\boldsymbol{\lambda}}^{pa}$ is constructed. Substituting $\hat{\boldsymbol{\lambda}}^{pa}$ into equation (6) yields $\hat{\boldsymbol{\beta}}^{pa}$, which are estimated population-averaged regression coefficients based on a model with multiple random effects.

### 3.1. *Marginalized Standard Errors*

The delta method can be used to obtain standard errors associated with the estimates $\hat{\boldsymbol{\beta}}^{pa}$. For this, let $\hat{\boldsymbol{\eta}}$ represent the $(p + r(r+1)/2) \times 1$ stacked vector of estimates $\hat{\boldsymbol{\beta}}^{ss}$ and $\hat{\boldsymbol{T}}$. To simplify the notation, let $p^* = (p + r(r+1)/2)$, and denote the estimated variance-covariance matrix of the $p^*$ parameter estimates as $V(\hat{\boldsymbol{\eta}})$. Then, using the delta method,

$$V(\hat{\boldsymbol{\beta}}^{pa}) = \frac{\partial\hat{\boldsymbol{\beta}}^{pa}}{\partial\hat{\boldsymbol{\eta}}'}\, V(\hat{\boldsymbol{\eta}})\, \frac{\partial(\hat{\boldsymbol{\beta}}^{pa})'}{\partial\hat{\boldsymbol{\eta}}} \ . \tag{7}$$

Here, $\frac{\partial\hat{\boldsymbol{\beta}}^{pa}}{\partial\hat{\boldsymbol{\eta}}'}$ is a matrix of size $p \times p^*$, $V(\hat{\boldsymbol{\eta}})$ is of size $p^* \times p^*$, and $\frac{\partial(\hat{\boldsymbol{\beta}}^{pa})'}{\partial\hat{\boldsymbol{\eta}}}$ is of size $p^* \times p$. Using the chain rule,

$$\frac{\partial\hat{\boldsymbol{\beta}}^{pa}}{\partial\hat{\boldsymbol{\eta}}'} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\frac{\partial\hat{\boldsymbol{\lambda}}^{pa}}{\partial\hat{\boldsymbol{\eta}}'} \ , \tag{8}$$

where

$$\frac{\partial\hat{\boldsymbol{\lambda}}^{pa}}{\partial\hat{\boldsymbol{\eta}}'} = \frac{\sum_{q=1}^{Q^r} \Psi(\hat{\lambda}_{ijq})(1 - \Psi(\hat{\lambda}_{ijq}))\ A(\boldsymbol{B}_q)\ \left(\partial\hat{\lambda}_{ijq}/\partial\hat{\boldsymbol{\eta}}'\right)}{\hat{\boldsymbol{\pi}}^{pa}(1 - \hat{\boldsymbol{\pi}}^{pa})} \tag{9}$$

with $\hat{\lambda}_{ijq} = \boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}}^{ss} + \boldsymbol{z}'_{ij}\hat{\boldsymbol{T}}\boldsymbol{B}_q$. In the above equation, the numerator is a $1 \times p^*$ row vector that is obtained for each of the $n$ observations, which upon stacking yields a $n \times p^*$ matrix. The denominator is the elementwise product of the two $n \times 1$ vectors. Thus, to form this derivative matrix, each element in a row of the $n \times p^*$ matrix in the numerator is divided by the corresponding element of the $n \times 1$ denominator vector. Finally, the $1 \times p^*$ row vector is given as

$$\frac{\partial\hat{\lambda}_{ijq}}{\partial\hat{\boldsymbol{\eta}}'} = \left[\ \boldsymbol{x}'_{ij} \vdots (\boldsymbol{B}'_q \otimes \boldsymbol{z}'_{ij})\mathsf{J}'_r\ \right] \ , \tag{10}$$

where $\otimes$ represents the Kronecker product, and $\mathsf{J}_r$ is the $(r \times (r+1)/2) \times r^2$ elimination matrix, which removes the elements above the main diagonal of a lower triangular matrix (Magnus, 1988).

### 3.2. *Other Link Functions*

The procedure can be extended to other link functions as well. For example, for a probit mixed model, population-averaged probabilities are obtained as:

$$\hat{\pi}_{ij}^{pa} \approx \sum_{q=1}^{Q^r} \Phi(\boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}}^{ss} + \boldsymbol{z}'_{ij}\hat{\boldsymbol{T}}\boldsymbol{B}_q)A(\boldsymbol{B}_q) \ , \tag{11}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution, and regression coefficients are obtained as:

$$\hat{\boldsymbol{\beta}}^{pa} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ \Phi^{-1}\left(\hat{\boldsymbol{\pi}}^{pa}\right) \ , \tag{12}$$

with $\Phi^{-1}(\cdot)$ as the probit function (inverse of the normal CDF). Standard errors are obtained using equation (7) with

$$\frac{\partial\hat{\boldsymbol{\beta}}^{pa}}{\partial\hat{\boldsymbol{\eta}}'} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ \frac{\sum_{q=1}^{Q^r} \phi(\hat{\lambda}_{ijq})\ A(\boldsymbol{B}_q)\ \left(\partial\hat{\lambda}_{ijq}/\partial\hat{\boldsymbol{\eta}}'\right)}{\phi\left(\Phi^{-1}(\hat{\boldsymbol{\pi}}^{pa})\right)} \ , \tag{13}$$

with $\phi(\cdot)$ is the PDF of the standard normal distribution, and again using equation (10) for the derivatives $\left(\partial\hat{\lambda}_{ijq}/\partial\hat{\boldsymbol{\eta}}'\right)$.

More generally, denoting $F(\cdot)$ and $F^{-1}(\cdot)$ as the CDF and the link function (inverse CDF), the following general forms are:

$$\hat{\pi}_{ij}^{pa} \approx \sum_{q=1}^{Q^r} F(\boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}}^{ss} + \boldsymbol{z}'_{ij}\hat{\boldsymbol{T}}\boldsymbol{B}_q)A(\boldsymbol{B}_q) \ , \tag{14}$$

$$\hat{\boldsymbol{\beta}}^{pa} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ F^{-1}\left(\hat{\boldsymbol{\pi}}^{pa}\right) \ , \tag{15}$$

$$\frac{\partial\hat{\boldsymbol{\beta}}^{pa}}{\partial\hat{\boldsymbol{\eta}}'} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ \ \partial F^{-1}\left(\hat{\boldsymbol{\pi}}^{pa}\right)\ \sum_{q=1}^{Q^r} A(\boldsymbol{B}_q)\ \ \partial F(\hat{\lambda}_{ijq})\left(\partial\hat{\lambda}_{ijq}/\partial\hat{\boldsymbol{\eta}}'\right). \tag{16}$$

The specific elements for several link functions are listed in Table 1. It should be noted that the loglog link is sometimes presented as the negative loglog link (i.e., $-\log(-\log(\pi))$), in which case the elements would be of opposite sign.

## 4. Simulation Studies

To evaluate and compare the performance of the proposed marginalization method, two small-scale simulation studies

**Table 1**
*Corresponding elements for several link functions, where $F$ = CDF, $\pi = Pr(Y = 1)$, and $\lambda = \boldsymbol{x}'\boldsymbol{\beta}^{ss} + \boldsymbol{z}'\boldsymbol{TB}$*

| Link | $F^{-1}$ | $F$ | $\partial F^{-1}$ | $\partial F$ |
|---|---|---|---|---|
| Logit | $\log(\pi/(1-\pi))$ | $1\ /\ (1+\exp(-\lambda))$ | $1\ /\ (\pi(1-\pi))$ | $F(\lambda)(1-F(\lambda))$ |
| Probit | $\Phi^{-1}(\pi)$ | $\Phi(\lambda)$ | $1/(\phi(\Phi^{-1}(\pi)))$ | $\phi(\lambda)$ |
| Cloglog | $\log(-\log(1-\pi))$ | $1-\exp(-\exp(\lambda))$ | $1/((\pi-1)\log(1-\pi))$ | $(F(\lambda)-1)(-\exp(\lambda))$ |
| Loglog | $\log(-\log(\pi))$ | $\exp(-\exp(\lambda))$ | $1/(\pi\log(\pi))$ | $F(\lambda)(-\exp(\lambda))$ |

were conducted. The first considered a random-intercept probit model in which marginalized estimates can be expressed in closed form. This allows us to compare our approach to the marginalized (closed form) estimates of the random-intercept probit model. The second simulation study considered a logistic mixed model with random intercepts and time trends. Here, we compared results of the proposed method with estimates from GEE and marginalized multilevel models. We also investigated the role of MAR missing data on our proposed approach, and contrasted it with GEE estimates.

### 4.1. *Random-Intercept Probit Model*

The specifications for this simulation study were based on the example and data analysis presented in Section 5. To examine the effect of sample size, 1000 datasets were simulated each with 100 or 500 subjects and four repeated observations according to a marginal probit regression model

$$\Phi^{-1}\left(\pi_{ij}^{pa}\right) = \beta_0 + \beta_1 G_i + \beta_2 T_j + \beta_3 (G_i \times T_j)\ , \qquad (17)$$

where $G_i$ is a subject-level grouping variable coded 0 or 1 (half of the 100 or 500 subjects had $G_i = 0$, while the other half were coded as 1), $T_j$ represents time and is coded 0, 1, 2, 4 for the four timepoints, respectively. The values of the regression coefficients on the marginal scale (PA) were set as: $\beta_0 = -1.1259$, $\beta_1 = 0.4505$, $\beta_2 = 0.027$, and $\beta_3 = -0.1147$. These were selected to yield response proportions of .130, .136, .142, and .155 ($G_i = 0$) and .250, .223, .198, and .155 ($G_i = 1$), for the two groups and four timepoints, respectively. Additionally, the random intercept variance was set as $\sigma_\upsilon^2 = 1$ to reflect an intraclass correlation, $r = \sigma_\upsilon^2/(\sigma_\upsilon^2 + 1)$, of 0.5 for the repeated observations. Thus, the SS regression coefficients would be equal to the PA coefficients multiplied by a factor of $\sqrt{2}$.

To generate the binary data, four random normal variates were obtained for each of the 100 or 500 subjects with mean 0 and variance-covariance matrix $\sigma_\upsilon^2 \mathbf{1}\mathbf{1}' + \boldsymbol{I}$. These random variates were then dichotomized in accordance with the marginal response probabilities for the two groups across the four timepoints given in the preceding paragraph.

The closed form marginal estimates of the random intercept probit model were obtained using SAS PROC NLMIXED, while the proposed marginalization parameters were estimated with Supermix (Hedeker et al., 2008) based on a random intercept probit model. For both, adaptive 7-point Gauss–Hermite quadrature was used. Table 2 presents the simulation results for both sample sizes of 100 and 500 subjects. In Table 2, bias is the difference between the average

**Table 2**
*Random-intercept probit model simulation results (1000 datasets with either N=100 or N=500). CF: closed form marginalization; PM: proposed marginalization.*

| $N$ | Model | Parameter | True value | Bias | Coverage | RMSE | Width |
|---|---|---|---|---|---|---|---|
| 100 | CF | Intercept | −1.1259 | −0.0157 | 0.951 | 0.1940 | 0.7385 |
| | | $T_j$ | 0.0270 | 0.0019 | 0.950 | 0.0651 | 0.2530 |
| | | $G_i$ | 0.4505 | 0.0093 | 0.947 | 0.2556 | 0.9832 |
| | | $G_i \times T_j$ | −0.1147 | −0.0030 | 0.956 | 0.0891 | 0.3465 |
| | PM | Intercept | −1.1259 | −0.0150 | 0.950 | 0.1944 | 0.7413 |
| | | $T_j$ | 0.0270 | 0.0019 | 0.951 | 0.0651 | 0.2529 |
| | | $G_i$ | 0.4505 | 0.0088 | 0.949 | 0.2556 | 0.9856 |
| | | $G_i \times T_j$ | −0.1147 | −0.0028 | 0.957 | 0.0891 | 0.3462 |
| 500 | CF | Intercept | −1.1259 | −0.0052 | 0.952 | 0.0839 | 0.3276 |
| | | $T_j$ | 0.0270 | −0.0002 | 0.951 | 0.0286 | 0.1113 |
| | | $G_i$ | 0.4505 | −0.0022 | 0.960 | 0.1098 | 0.4368 |
| | | $G_i \times T_j$ | −0.1147 | 0.0021 | 0.941 | 0.0393 | 0.1523 |
| | PM | Intercept | −1.1259 | −0.0048 | 0.952 | 0.0839 | 0.3280 |
| | | $T_j$ | 0.0270 | −0.0002 | 0.952 | 0.0285 | 0.1111 |
| | | $G_i$ | 0.4505 | −0.0027 | 0.960 | 0.1097 | 0.4371 |
| | | $G_i \times T_j$ | −0.1147 | 0.0023 | 0.941 | 0.0393 | 0.1521 |

parameter estimate and the true value, coverage is the proportion of 1000 datasets in which the 95% confidence interval included the true value, width is the average width of the 95% confidence interaval, and root mean squared error (RMSE) for a given parameter $\theta$ equals $\sqrt{E[(\hat{\theta}-\theta)^2]}$. As can be seen, for both sample sizes of 100 and 500, the proposed marginalization method provides near-identical results to the closed form marginal probit estimates. Both approaches recover the parameter values well with low bias and decent coverage.

### 4.2. *Random-Intercept and Trend Logistic Model*

The specifications for this simulation study were also based on the example and data analysis presented in Section 5. Here, 1000 datasets were simulated, each with 500 subjects and four repeated observations according to a marginal logistic regression model

$$\log\left[\frac{\pi_{ij}^{pa}}{1-\pi_{ij}^{pa}}\right] = \beta_0 + \beta_1 G_i + \beta_2 T_j + \beta_3 (G_i \times T_j)\ , \qquad (18)$$

where $G_i$ and $T_j$ were as previously defined. Again, the marginal regression coefficients were determined based on the same response proportions for the two groups across the four timepoints. Here, on the logistic scale, they were: $\beta_0 = -1.9$, $\beta_1 = 0.8$, $\beta_2 = 0.05$, and $\beta_3 = -0.2$. Additionally, the variance-covariance structure for the mixed model allowed for random intercepts ($\sigma_I^2 = 3$), time-trends ($\sigma_T^2 = 1$), and covariance ($\sigma_{IT} = 0.08$); these values were based on and similar to the variance covariance estimates for the example in Section 5. Similar to the probit simulation, four random normal variates were obtained for each of the 500 subjects with mean 0 and variance-covariance matrix $\boldsymbol{Z}\boldsymbol{\Sigma}_{\upsilon}\boldsymbol{Z}' + \pi^2/3\boldsymbol{I}$, which were then dichotomized according to the same marginal response probabilities for the two groups across the four timepoints.

Table 3 presents results from the following: the GEE approach using an unstructured working correlation matrix, the proposed marginalization method, and a marginalized logistic-probit-normal model as described by Griswold and Zeger (2004) including random subject intercepts and trends. In this latter model, the regression coefficients are on the marginal logistic scale, while the random effects variance-covariance parameters are on the probit scale. As Griswold and Zeger (2004) note, the probit is a convenient choice for the variance-covariance parameters for computational simplicity (i.e., to avoid an additional numerical integration). The parameters of the GEE and marginalized multilevel models were estimated in SAS (PROCs GENMOD and NLMIXED, respectively), while the proposed marginalization parameters were estimated with Supermix based on a random intercept and trend logistic model. For both the marginalized multilevel model and the mixed logistic model, adaptive 7-point Gauss–Hermite quadrature was used. For the GEE, robust standard errors were used to obtain the coverage probabilities.

As can be seen from Table 3, the results from all three models are near-identical, with low levels of bias and good coverage levels (i.e., near the nominal 95% level). As one would expect, the widths are appreciably smaller for the time-related parameters than for the intercept-related parameters. Finally, RMSE levels are relatively small, especially for the time-related parameters.

To examine the effect of missing data on the proposed approach, we generated data under a MAR mechanism. Using the same data generation procedure described above, we additionally imposed the following specifications. For a $G_i = 0$ (or $G_i = 1$) observation, if their first outcome was 0 (or 1), they had a 60 (or 80%) chance of dropping out (i.e., being missing at the next three timepoints). Similarly, for a $G_i = 0$ (or $G_i = 1$) observation, if their second outcome was 0 (or 1), they had a 80 (or 85%) chance of dropping out (i.e., being missing at the next two timepoints). Finally, for a $G_i = 0$ (or $G_i = 1$) observation, if their third outcome was 0 (or 1), they had a 90% chance of being missing at the last timepoint. Thus, the two groups were treated in an opposite manner based on their observed outcomes, and the amount of missing data was rather large. The observed proportions were .129, .170, .216, and .298 ($G_i = 0$) and .251, .176, .123, and .067 ($G_i = 1$), for the two groups and four timepoints, respectively, indicating

**Table 3**
*Logistic regression simulation results (1000 datasets with N=500). GEE UN: GEE with unstructured working correlation structure; MM: marginalized multilevel model with random intercept and time effects; PM: proposed marginalization from a mixed model with random intercepts and time effects.*

| Model | Parameter | True value | Bias | Coverage | RMSE | Width |
|---|---|---|---|---|---|---|
| GEE UN | Intercept | −1.90 | −0.0032 | 0.949 | 0.152 | 0.606 |
| | $G_i$ | 0.80 | 0.0002 | 0.944 | 0.204 | 0.781 |
| | $T_j$ | 0.05 | −0.0033 | 0.950 | 0.056 | 0.214 |
| | $G_i \times T_j$ | −0.20 | 0.0007 | 0.947 | 0.077 | 0.292 |
| MM | Intercept | −1.90 | 0.0014 | 0.947 | 0.156 | 0.605 |
| | $G_i$ | 0.80 | −0.0034 | 0.942 | 0.203 | 0.777 |
| | $T_j$ | 0.05 | −0.0078 | 0.933 | 0.057 | 0.216 |
| | $G_i \times T_j$ | −0.20 | 0.0030 | 0.953 | 0.077 | 0.291 |
| PM | Intercept | −1.90 | −0.0055 | 0.956 | 0.148 | 0.595 |
| | $G_i$ | 0.80 | −0.0112 | 0.946 | 0.200 | 0.770 |
| | $T_j$ | 0.05 | −0.0062 | 0.947 | 0.056 | 0.215 |
| | $G_i \times T_j$ | −0.20 | 0.0096 | 0.944 | 0.076 | 0.288 |

**Table 4**
*Logistic regression simulation results (1000 datasets with N=500) with missing data. GEE UN: GEE with unstructured working correlation structure; PM: proposed marginalization from a mixed model with random intercepts and time effects.*

| Model | Parameter | True value | Bias | Coverage | RMSE | Width |
|---|---|---|---|---|---|---|
| GEE UN | Intercept | −1.90 | −0.0141 | 0.9481 | 0.1889 | 0.7335 |
| | $G_i$ | 0.80 | −0.0155 | 0.9481 | 0.2380 | 0.8865 |
| | $T_j$ | 0.05 | −0.0576 | 0.9888 | 0.1544 | 0.6245 |
| | $G_i \times T_j$ | −0.20 | 0.1303 | 0.8576 | 0.2518 | 0.6507 |
| PM | Intercept | −1.90 | −0.0121 | 0.9500 | 0.1767 | 0.6745 |
| | $G_i$ | 0.80 | −0.0089 | 0.9490 | 0.2270 | 0.8662 |
| | $T_j$ | 0.05 | −0.0030 | 0.9540 | 0.1133 | 0.4437 |
| | $G_i \times T_j$ | −0.20 | 0.0115 | 0.9340 | 0.1573 | 0.5952 |

very different trends across time for the two groups, relative to the complete data scenario. The retention rates for the two groups were 100, 48, 25, and 14% ($G_i = 0$) and 100, 80, 69, and 63% ($G_i = 1$) for the two groups and four timepoints, respectively. Thus, both the reasons and amount of missing data varied considerably for the two groups.

Table 4 lists the simulation results for the GEE and proposed marginalization approach. As can be seen, the time-related parameters are rather poorly estimated by GEE, and the coverage rates are off as well, especially for the group by time interaction. Conversely, the proposed marginalization approach does reasonably well with low bias and close to nominal coverage rates. For both, the widths are appreciably larger than their counterparts under complete data, reflecting the effect of missing data. RMSE values are also larger than under the complete data scenario, and tend to favor the proposed marginalization approach.

## 5. Example

As mentioned in Section 1, Gruder et al. (1993) present data from a smoking-cessation study in which 489 subjects were assessed for their smoking status (y/n) at four timepoints: post-intervention, and 6, 12, and 24 months later. Subjects were randomized to either a control, discussion, or social support condition. Here, for simplicity, subjects in the two treatment conditions of discussion and social support are combined into one group. The missing data pattern across time was monotonic, in that subjects who were missed at a given follow-up were also missing thereafter. The number of subjects measured at the four timepoints was 489, 454, 429, and 372, respectively. Smoking abstinence rates for the two groups across time are presented in Table 5. These show that the control rates are somewhat constant across time, while the treatment group rates decline after the end-of-program timepoint. Also included in Table 5, are the observed correlations of the smoking abstinence outcomes across time. These pairwise correlations are clearly not equal, which would violate the assumption of a random-intercept model.

The outcome $Y_{ij}$ was coded 0=smoking and 1=quit, and several logistic regression models were applied to these data with covariates for time, group, and group by time interaction. The time variable was coded so that a unit represents 6 months ($T_j = 0, 1, 2, 4$), and the group variable was an indicator of treatment ($G_i = 0$ for control, and $G_i = 1$ for treatment). Table 6 presents estimates from the following: the GEE approach using an independent working correlation matrix (GEE IND) and an unstructured working correlation matrix (GEE UN), a mixed effects model with random subject intercepts and time trends (MRM), a marginalized logistic-probit-normal model including random subject intercepts and trends (MM), and the proposed marginalization approach (PM).

As can be seen in Table 6, the subject-specific MRM estimates are not on the same numerical scale as the population-averaged estimates of the GEE. However, the proposed marginalization of these MRM estimates yields results that are in close agreement with those obtained from both the GEE and the marginalized logistic-probit-normal models. An advantage for the user of the proposed marginalization is that both subject-specific and population-averaged estimates and inferences can be obtained from the same analysis. In the present example, both approaches indicate a significant

**Table 5**
*Point prevalence rates and correlations of smoking abstinence over time.*

| | Post-intervention | 6 months | 12 months | 24 months |
|---|---|---|---|---|
| Control group ($N = 109, 97, 92, 77$) | 0.174 | 0.072 | 0.185 | 0.182 |
| Treatment group ($N = 380, 357, 337, 295$) | 0.345 | 0.182 | 0.196 | 0.217 |
| Correlations | | | | |
| Post-intervention | 1.000 | 0.332 | 0.287 | 0.263 |
| 6 months | 0.332 | 1.000 | 0.481 | 0.337 |
| 12 months | 0.287 | 0.481 | 1.000 | 0.493 |
| 24 months | 0.263 | 0.337 | 0.493 | 1.000 |

**Table 6**
*Logistic regression estimates (standard errors), and p-values. GEE IND: GEE with independent working correlation structure; GEE UN: GEE with unstructured working correlation structure; MRM: Mixed model with random intercept and time effects; MM: Marginalized multilevel model with random intercept and time effects; PM: Proposed marginalized estimates from MRM.*

| Model | Intercept | $T$ | $G$ | $G \times T$ |
|---|---|---|---|---|
| GEE IND | −1.839 | 0.074 | 0.902 | −0.221 |
| | (0.256) | (0.094) | (0.275) | (0.103) |
| | 0.001 | 0.431 | 0.001 | 0.032 |
| GEE UN | −1.841 | 0.075 | 0.905 | −0.201 |
| | (0.256) | (0.092) | (0.275) | (0.101) |
| | 0.001 | 0.418 | 0.001 | 0.047 |
| MRM | −2.785 | −0.479 | 1.478 | −0.331 |
| | (0.430) | (0.270) | (0.410) | (0.247) |
| | 0.001 | 0.076 | 0.001 | 0.180 |
| MM | −1.889 | 0.109 | 0.941 | −0.211 |
| | (0.240) | (0.086) | (0.260) | (0.095) |
| | 0.001 | 0.206 | 0.001 | 0.027 |
| PM | −1.885 | 0.092 | 0.945 | −0.232 |
| | (0.236) | (0.093) | (0.257) | (0.102) |
| | 0.001 | 0.321 | 0.001 | 0.023 |

group effect such that there is a benefit for treatment at post-intervention (e.g., from GEE UN, $z = 0.905/0.275 = 3.29$). However, the marginal models indicate a significant group by time interaction such that the treatment benefit diminishes across time. Interestingly, the MRM does not yield a significant group by time interaction. Thus, the conclusions are different depending on whether one controls for the subject effects or not. Such situations in which the conclusions vary between marginal and conditional approaches have been previously reported in the literature (Agresti, 1989; Lindsey and Lambert, 1998).

In Table 6, we have only presented the regression coefficient estimates, but the random intercept and trend variance estimates of the MRM were large ($\hat{\sigma}_I^2 = 3.774$, standard error =1.214, and $\hat{\sigma}_T^2 = 1.302$, standard error = 0.451) reflecting a good deal of subject heterogeneity in the data. The covariance between the two random effects was small ($\hat{\sigma}_{IT} = 0.170$, standard error = 0.359) indicating that a subject's post-intervention smoking status was not related to their trend across time.

## 6. Discussion

In this article, we have presented a simple approach using numerical quadrature to produce marginal estimates and standard errors for the regression coefficients from mixed model estimates. Specifically, we have presented results for several link functions applied to binary outcome data. We have also derived results (not shown) for proportional and non-proportional odds models of ordinal data under all of the presented link functions, as well as for nominal outcomes under the logistic link. An advantage of the proposed approach is that both subject-specific and population-averaged estimates (and standard errors) can be obtained from one analysis with minimal additional computation. In addition, although we have described our approach in terms of two-level (longitudinal) data, extending to three or more levels is relatively straightforward.

The limited simulation studies confirmed that for a random intercept model the marginalized estimates were near-identical to closed form marginal probit estimates. Extending to a random intercept and trend logistic model, the marginalized estimates also agreed quite closely to marginalized multilevel and GEE estimates. A further simulation allowing for missing data, under MAR missingness, showed that the GEE estimates of the time-related parameters were poorly estimated, whereas the proposed approach produced reasonable results. This agrees with the theory that indicates that mixed models using full likelihood estimation provide valid results under MAR, whereas GEE only does so under MCAR. It should be emphasized that the amount of missing data in the simulation was rather large and strictly followed an MAR mechansim. In addition, extensions to allow MAR missingness under the GEE approach have been developed based on weighting (Robins et al., 1995) and imputation (Paik, 1997). Nonetheless, the proposed approach yields reasonable marginal estimates under MAR missingness with rather minimal effort. An important caveat is that the variance-covariance structure of the data must be properly specified in a mixed model or biased results can be obtained (Hedeker and Gibbons, 2006; Kwok et al., 2007).

In the example presented, we found close agreement of our marginalized estimates to those obtained from marginalized multilevel models. This is not surprising as our approach can be viewed as essentially an alternative way of producing marginalized multilevel estimates. However, given that software for marginalized multilevel models is somewhat limited, whereas software for mixed models is widely available, our approach can be used in conjunction with mixed model estimates from the many different programs that exist. The marginalized results presented in this article were obtained using Supermix (Hedeker et al., 2008), which produces both SS and PA estimates automatically.

## 7. Supplementary Materials

The dataset and a SAS script using PROCs NLMIXED and IML that can be used to obtain the marginalized estimates are available at the *Biometrics* website on Wiley Online Library. These materials are also available from the first author upon request.

### Acknowledgements

This work was supported by National Heart, Lung, and Blood Institute grant R01 HL121330 (Hedeker). The authors thank Robin Mermelstein for use of the smoking cessation dataset. The authors also thank the Editor, Associate Editor, and Referee for several excellent points raised in the review of this article which greatly improved it.

## References


Agresti, A. (1989). A survey of models for repeated ordered categorical response data. *Statistics in Medicine* **8**, 1209–1224.

Agresti, A. (2002). *Categorical Data Analysis, 2nd edition*. New York: Wiley.

Brown, H. and Prescott, R. (2006). *Applied Mixed Models in Medicine, 2nd edition*. Chichester, UK: Wiley.

Demidenko, E. (2004). *Mixed Models Theory and Applications*. Hoboken, NJ: Wiley.

Demirtas, H. (2004). Assessment of relative improvement due to weights within generalized estimating equations framework for incomplete clinical trials data. *Journal of Biopharmaceutical Statistics* **14**, 1085–1098.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2011). *Applied Longitudinal Analysis, 2nd edition*. Hoboken, NJ: Wiley.

Goldstein, H. (2011). *Multilevel Statistical Models, 4th edition*. Chichester, UK: Wiley.

Griswold, M. E., Swihart, B. J., Caffo, B. S., and Zeger, S. L. (2013). Practical marginalized multilevel models. *Stat* **2**, 129–142.

Griswold, M. E. and Zeger, S. L. (2004). On marginalized multilevel models and their computation. John Hopkins University, Department of Biostatistics Working Papers, Working Paper 99.

Gruder, C. L., Mermelstein, R. J., Kirkendol, S., Hedeker, D., Wong, S. C., Schreckengost, J., et al. (1993). Effects of social support and relapse prevention training as adjuncts to a televised smoking cessation intervention. *Journal of Consulting and Clinical Psychology* **61**, 113–120.

Heagerty, P. J. and Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science* **15**, 1–26.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal Data Analysis*. Hoboken, NJ: Wiley.

Hedeker, D., Gibbons, R. D., du Toit, M., and Cheng, Y. (2008). *SuperMix: Mixed Effects Models*. Lincolnwood, IL: Scientific Software International.

Kwok, O., West, S. G., and Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research* **42**, 557–592.

Lindsey, J. K. and Lambert, P. (1998). On the appropriateness of marginal models for repeated measurements in clinical trials. *Statistics in Medicine* **17**, 447–469.

Magnus, J. R. (1988). *Linear Structures*. New York: Oxford University Press.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models, 2nd edition*. Hoboken, NJ: Wiley.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York, NY: Springer-Verlag.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* **59**, 25–35.

Paik, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association* **92**, 1320–1329.

Preisser, J. S., Lohman, K. K., and Rathouz, P. J. (2002). Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine* **21**, 3035–3054.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis, 2nd edition*. Thousand Oaks, CA: Sage.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semi-parametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling*. Boca Raton, FL: Chapman and Hall/CRC.

Stroud, A. H. and Sechrest, D. (1966). *Gaussian Quadrature Formulas*. Englewood Cliffs, NJ: Prentice Hall.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–130.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**, 1049–1060.


*Received February* 2017. *Revised March* 2017.
*Accepted March* 2017.