Title

Covariate Decomposition Methods for Longitudinal Missing-at-Random Data and Predictors Associated with Subject-Specific Effects

Journal

Australian & New Zealand Journal of Statistics, 56(4)

ISSN

1369-1473

Authors

Neuhaus, John M
McCulloch, Charles E

Publication Date

2014-12-01

DOI

10.1111/anzs.12093

Peer reviewed

HHS Public Access

Author manuscript

Aust N Z J Stat. Author manuscript; available in PMC 2015 June 04.

Published in final edited form as:

Aust N Z J Stat. 2014 December ; 56(4): 331–345. doi:10.1111/anzs.12093.

COVARIATE DECOMPOSITION METHODS FOR LONGITUDINAL MISSING-AT-RANDOM DATA AND PREDICTORS ASSOCIATED WITH SUBJECT-SPECIFIC EFFECTS

John M. Neuhaus[*] and **Charles E. McCulloch**

University of California, San Francisco

Summary

Investigators often gather longitudinal data to assess changes in responses over time within subjects and to relate these changes to within-subject changes in predictors. Missing data are common in such studies and predictors can be correlated with subject-specific effects. Maximum likelihood methods for generalized linear mixed models provide consistent estimates when the data are `missing at random' (MAR) but can produce inconsistent estimates in settings where the random effects are correlated with one of the predictors. On the other hand, conditional maximum likelihood methods (and closely related maximum likelihood methods that partition covariates into between- and within-cluster components) provide consistent estimation when random effects are correlated with predictors but can produce inconsistent covariate effect estimates when data are MAR. Using theory, simulation studies, and fits to example data this paper shows that decomposition methods using complete covariate information produce consistent estimates. In some practical cases these methods, that ostensibly require complete covariate information, actually only involve the observed covariates. These results offer an easy-to-use approach to simultaneously protect against bias from both cluster-level confounding and MAR missingness in assessments of change.

Keywords

bias; conditional likelihood; confounding; consistent estimation

1. Introduction

1.1. Change as the scientific objective

Investigators gather longitudinal and clustered data to assess changes in responses over time or within clusters and to relate these changes to within-subject or -cluster changes in predictors. For example, in the late 1990s, Haan *et al*. (2003) assembled a cohort of 1735 community-dwelling Mexican-Americans who were 60–98 years old and lived in the Sacramento, California, area. The study, known as the Sacramento Area Latino Study on Aging (SALSA), assessed the subjects' physical and cognitive functioning every 12–15

[*]Author to whom correspondence should be addressed. Department of Epidemiology and Biostatistics, University of California, San Francisco, California 94143-0560, USA. john@biostat.ucsf.edu.

months for up to seven study visits. The objective of the longitudinal study was to evaluate the effects of metabolic and cardiovascular risk factors on changes in cognitive functioning and the incidence of dementia. The study evaluated several measures of cognitive functioning but this paper will focus on a measure of global cognitive ability known as the Modified Mini-Mental State Examination (3MSE). The 3MSE evaluates memory, orientation, attention and language on a scale of 0 to 100. Investigators also use a binary version which cuts the 3MSE score at the 20th percentile or 80. The investigators were interested in assessing rates of change in cognitive functioning and relating these changes to changes in metabolic and cardiovascular risk factors. One particular question of interest is to assess whether maintenance of physical functioning protects against cognitive decline.

Figure 1 displays subject-specific trajectories of 3MSE as a function of age for 12 subjects selected to illustrate the wide variety of patterns. Figure 1 identifies several features that statistical models need to address. First, the subjects display variability in initial and average levels of 3MSE, as well as differences in changes with age. Second, not all the subjects provided data at all the seven study visits. Examination of data availability patterns in SALSA shows that many subjects dropped out before the end of the study. As Figure 2 indicates, 1735 subjects provided data at the first visit but only 766 subjects were measured at visit seven. Subject-specific variability in the levels of the outcome and rates of change over time, as well as drop-out are standard features of longitudinal studies.

A typical approach to assessing the association of within-subject changes in physical functioning with changes in cognitive functioning would be to fit a generalized linear mixed model (McCulloch, Searle & Neuhaus 2008) which accommodates dependence of responses within subjects by including random effects. In particular, one might fit a mixed-effects logistic model to the binary repeated measures of poor cognitive functioning (3MSE < 80) using the time varying values of a physical functioning measure of interest in SALSA known as Activities of Daily Living (ADL) as the predictor of interest. The ADL scale is a count of the number of activities such as ability to feed, bathe, dress, and groom oneself that the subject could not perform and the SALSA study assessed this measure for each subject at every visit. Denoting the *j*th ADL measurement of the *i*th subject by ADL_{ij}, a standard mixed-effects logistic model would include ADL_{ij} and at least subject-specific intercepts to allow subject-specific variability in the probability of poor cognitive functioning. Rather than integrating the subject-specific intercepts out of the likelihood, as mixed-effects models do, analysts often prefer to use a conditional likelihood approach (McCulloch *et al*. 2008) which removes the intercepts from the analysis using the likelihood conditional on the sufficient statistics for the intercepts. A conditional likelihood approach depends on ADL_{ij} only through the deviations of each measurement from the subject-specific mean, $ADL_{ij} - \overline{ADL}_i$ (Neuhaus & McCulloch 2006), where $\overline{ADL}_i$ is the mean for the *i*th subject. This form of predictor dependence suggests a third analytic approach where the analyst decomposes ADL_{ij} into a within-subject component, $ADL_{ij} - \overline{ADL}_i$, as well as a between-subject component, $\overline{ADL}_i$ and includes both components in a mixed-effects logistic model with separate regression coefficients β_W and β_B attached to the two components (Neuhaus & Kalbfleisch 1998; Neuhaus & McCulloch 2006). Rather than decompose covariates based on a subject-specific mean, an analyst may prefer to decompose covariates based on the

initial or baseline measurement, ADL_{i1} and thus focus on $ADL_{ij} - ADL_{i1}$. This fourth approach assesses the association of subject-specific changes from baseline in ADL with poor cognitive functioning. In addition to ADL, each of the above four fitting approaches also included age as a predictor, using the same decompositions into between- and within-subject components as we used with ADL.

Table 1 presents the estimated covariate effects and associated standard errors (as subscripts) for each of these four approaches, where the superscript *o* indicates deviations compared to the observed mean. We will discuss these results in more detail in Section 4 but note here that the four approaches provide very different estimates of the associations of ADL and age with poor cognitive functioning. First, estimates of the between- and within-subject associations, $\hat{\beta}_B$ and $\hat{\beta}_W$, respectively, are very different for both ADL and age suggesting associations between the covariates and the subject-specific effects. Estimates based on the two covariate decomposition methods are also different for both ADL and age. In addition, the estimated within-subject associations from the conditional likelihood and the two covariate decomposition approaches differ substantially from the estimates based on a standard mixed-effects logistic model. Finally, the conditional likelihood estimates and the estimates from the covariate decomposition approach based on the observed subject-specific means, $\overline{ADL_i}$ and $\overline{AGE_i}$, are nearly identical, as suggested above. One objective of this paper is to explain why the different approaches produce different estimates of the associations of within-subject changes in predictors with changes in response in longitudinal and clustered data studies.

1.2. Generalized linear mixed models

The class of generalized linear mixed models (GLMM), which extend the class of generalized linear models by adding random effects to the linear predictor, accommodates dependence of responses within subjects, subject-specific trajectories and varying numbers of responses for each subject that can be unequally spaced in time. To construct this class, we assume that the data of interest consist of longitudinal responses Y_{ij} along with *p*-dimensional covariates $\mathbf{x}_{ij}$, where *i* indexes subjects (i=1,…,m) and *j* indexes units within subjects (j=1,…,n_i), and that we want to assess the association of changes in $\mathbf{x}$ with a known function of $E(Y)$. Generalized linear mixed models specify that, given a vector b_i of parameters specific to the *i*th subject, for the *j*th unit, the conditional density of Y_{ij} is of the form

$$f_Y\left(y_{ij}|b_i,\mathbf{x}_{ij}\right)=exp\left[\left\{y_{ij}\theta_{ij}-c\left(\theta_{ij}\right)\right\}\phi+d\left(y_{ij},\phi\right)\right], \quad (1)$$

where *c* and *d* are functions of known form, φ is a scale parameter and θ_{ij} depends on the covariates $\mathbf{x}_{ij}$, as well as the random effects b_i. In addition, one assumes that

$$\mu_{ij}=E\left(Y_{ij}|b_i,\mathbf{z}_{ij},\mathbf{x}_{ij}\right)=g^{-1}\left(\mathbf{x}_{ij}^{\top}\beta+\mathbf{z}_{ij}^{\top}b_i\right), \quad (2)$$

where $\mathbf{x}_{ij}^{\top}$ and $\mathbf{z}_{ij}^{\top}$ are the known covariate row vectors relating the fixed and random effects, respectively, to the conditional mean of the observations, *g* is a link function and μ_{ij} is a function of θ_{ij}. The function $\eta_{ij}=g\left(\mu_{ij}\right)=\mathbf{x}_{ij}^{\top}\beta+\mathbf{z}_{ij}^{\top}b_i$ is known as the linear predictor. Given

b_i, we assume that the responses $Y_{i1}, \ldots, Y_{in_i}$ are independent. We complete model construction by specifying that the random effects b_i vary over subjects according to a multivariate distribution G, often Gaussian, with parameters Σ. Consistent with Figure 1, analysts are often interested in fitting models to describe subject-specific underlying level and changes over time which generalized linear mixed models accommodate by including random intercepts and slopes, respectively. In this case, b_i is a two dimensional vector, as is $\mathbf{z}_{ij}^{\top}$.

We construct the likelihood for (β, Σ) by integrating the conditional density (1) of the responses Y over the distribution of the random effects b. The likelihood for generalized linear mixed models fitted to m independent subjects, with the ith subject containing n_i units, is

$$\mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f_Y(y_{ij}|b, \mathbf{x}_{ij})\, dG(b), \quad (3)$$

where $Y_{ij}|b$ follows a generalized linear model of form (1)–(2). One can obtain estimates of the model parameters by maximizing (3). One can also obtain model-based standard error estimates of estimated model parameters from the information matrix of the fitted likelihood.

1.3. Complications in practice

Generalized linear mixed models require that the covariates, $\mathbf{x}_{ij}$, be uncorrelated with the random effects, b_i, but in practice observations often exhibit non-zero correlations. One reason for this is that the random effects b_i may include omitted covariates w_i that are associated both with $\mathbf{x}_{ij}$ and the response Y_{ij}. For example, in the SALSA analyses of ADL with cognitive functioning, both ADL measurements and underlying levels of cognitive functioning b_i may be related to a common factor such as overall level of health, resulting in $cor(ADL_{ij}, b_i) \neq 0$. Figure 3 provides evidence of this non-zero correlation. Based on the fit of a linear mixed-effects model with 3MSE as the response and ADL as the predictor, Figure 3 displays a plot of time 1 ADL values versus predicted random intercepts, b_i, as best linear unbiased predictions (BLUPs), along with an added non-parametric locally weighted scatterplot smoothing (LOWESS) curve. Figure 3 displays a clear correlation of ADL values and predicted random intercepts. Neuhaus & McCulloch (2006) showed that fitting standard generalized linear mixed models in settings with correlations between covariates and random effects produces biased estimates of associations of covariates with the response.

Neuhaus & Kalbfleisch (1998) and Neuhaus & McCulloch (2006) noted that generalized linear mixed models that decompose covariates into between- and within-subject components provide two important features: (i) they protect against bias due to associations of covariates and subject-specific effects; and (ii) they focus attention on the associations of within-subject change in covariates with within-subject change in the response, the typical objective of scientific interest in longitudinal studies. Neuhaus & Kalbfleisch (1998) and Neuhaus & McCulloch (2006) also showed that covariate decomposition methods are

closely related to conditional likelihood approaches that remove random effects from the likelihood by conditioning on sufficient statistics.

Investigators find covariate decomposition methods appealing not only for their useful statistical properties but because they provide covariate effect estimates of scientific interest. For example, Enders (2013) was interested in assessing the association of student socio-economic status with performance in mathematics using a sample of students clustered within a sample of schools. By decomposing the socio-economic status variable into between- and within-school components, the investigators were able to assess both the association of overall school socio-economic status using school-specific mean values and the association of the difference between student-specific socio-economic status and the school average with the mathematics performance outcome. Both of these covariate effects are of scientific interest.

Consistent with Figure 2, it is well-known that missing values and drop-out are ubiquitous in longitudinal studies. The effect of missing values on statistical analyses depends on the reason observations are missing. Little & Rubin (2002) provided a useful hierarchy and showed that statistical methods based on full maximum likelihood often provide consistent estimation when observations are missing at random. Although conditional likelihood approaches produce estimates using likelihood maximization methods, Rathouz (2004) and Roy *et al*. (2006) showed that these approaches can yield inconsistent estimation for logistic and Poisson models, respectively. These authors also derived modifications to standard conditional likelihoods that produce consistent estimation. The modified conditional likelihoods require specialized software and only apply to logistic and Poisson models. One objective of this paper is to investigate whether the inconsistent estimation result carries over to the closely related covariate decomposition methods. While Rathouz (2004) and Roy *et al*. (2006) show that conditional likelihood methods may provide inconsistent estimation when observations are missing at random, Skrondal & Rabe-Hesketh (2014) showed that conditional likelihood methods provide consistent estimation in settings where missingness depends on current outcomes, that is, a setting where observations are not missing at random.

The overall objective of this paper is to examine the performance of methods to assess change in longitudinal studies that feature the commonly occurring complications of observations that are missing at random and correlations of predictors with random effects. In particular, Section 2 develops theory to show that decomposition methods using complete covariate information produce consistent estimates. In some practical cases these methods, that ostensibly require complete covariate information, actually only involve the observed covariates. In Section 3 simulation studies are used to further illustrate the results, while Section 4 presents illustrative analyses of data from the SALSA study. The paper concludes with a discussion in Section 5.

2. Theory

2.1. Covariate decomposition methods

Motivated by the form of covariate dependence in conditional likelihood approaches, Neuhaus & Kalbfleisch (1998) and Neuhaus & McCulloch (2006) proposed expanding the term βx_{ij} in (2) into $\beta_W \left(x_{ij} - \bar{x}_i \right) + \beta_B \bar{x}_i$, where $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$, the observed mean of the covariate x_{ij} in the ith subject. The parameter β_W measures association of within-subject change in X with change in response Y, typically the scientific objective of longitudinal studies. Covariate decomposition methods have several advantages over conditional likelihood approaches. First, they apply more generally than conditional likelihood approaches, to non-canonical link functions and models with multiple random effects. Covariate decomposition methods allow separate assessment of between- and within-subject covariate effects and analysts can use standard GLMM routines with decomposed covariates to implement the approach. Neuhaus & McCulloch (2006) showed that in settings with associations of covariates with subject-specific effects, and in the absence of missing data, the within-subject estimator, $\hat{\beta}_W$, consistently estimates β in (2). On the other hand, the between-subject estimator, $\hat{\beta}_B$ converges to a value β* which Neuhaus & McCulloch (2006) showed may not equal β in (2).

2.2. Estimation with missing values

Our objective is to assess the performance of estimation methods in settings where observations are missing at random. We assume that the study intends to gather T_i responses $Y_{i1},\ldots, Y_{iT_i}$ from the ith subject. We define T_i missing value indicators $R_{i1},\ldots,R_{iT_i}$ such that R_{ij} =1 if Y_{ij} is observed and R_{ij} =0 if Y_{ij} is missing. We can partition the complete measurements of the ith subject into $(Y_{i1}, \ldots, Y_{iT_i}) = \left(Y_i^{(o)}, Y_i^{(m)} \right)$, where the superscript (o)=denotes observed and (m) denotes missing. Given covariates X, observations are said to be missing at random (MAR) if

$$Pr\left(R|Y^{(o)}, Y^{(m)}, X \right) = Pr\left(R|Y^{(o)}, X \right), \quad (4)$$

where we drop subscripts i and j for notational convenience.

The standard argument for consistent estimation using maximum likelihood under MAR decomposes the observed data as follows, where $[\mathscr{A}|\mathscr{B}]$ denotes the conditional density of $\mathscr{A}$ given $\mathscr{B}$:

$$\begin{aligned} \left[Y^{(o)}, R|X \right] &= \int \left[Y^{(o)}, Y^{(m)}, R|X \right] dY^{(m)} \\ &= \int \left[R|Y^{(o)}, Y^{(m)}, X \right] \left[Y^{(o)}, Y^{(m)}|X \right] dY^{(m)} \\ &= \left[R|Y^{(o)}, X \right] \int \left[Y^{(o)}, Y^{(m)}|X \right] dY^{(m)} \\ &= \left[R|Y^{(o)}, X \right] \left[Y^{(o)}|X \right], \end{aligned} \quad (5)$$

where we obtain (5) under an MAR assumption (4) given X.

Assuming that $[R|Y^{(o)}, X]$ is free of β implies that we can obtain consistent estimation using just $[Y^{(o)}|X]$ and maximum likelihood. It is important to note that we construct models for Y based on X which provides $[Y^{(o)}|X]$ for `free'. This is because, once we have specified $[Y|X]$, we can obtain $[Y^{(o)}|X]$ simply by restricting attention to the subset of data that are observed. When X consists of all the between- and within-subject components resulting from covariate decomposition, the general theory above shows that $\hat{\beta}_W$ and $\hat{\beta}_B$ converge to the same values as in the no missing response case. In particular, in settings with associations of covariates with subject-specific effects and with responses Y missing at random (MAR), $\hat{\beta}_W$ consistently estimates β in (2), whereas $\hat{\beta}_B$ may not.

Now we consider settings, such as subject drop-out, where we may not observe the covariates corresponding to missing responses, Y. Mechanically, the same argument applies when we work with distributions conditional on the observed covariates $X^{(o)}$. In this case, we decompose the observed data as:

$$\begin{aligned}\left[Y^{(o)}, R|X^{(o)}\right] &= \int \left[Y^{(o)}, Y^{(m)}, R|X^{(o)}\right] dY^{(m)} \\ &= \int \left[R|Y^{(o)}, Y^{(m)}, X^{(o)}\right] \left[Y^{(o)}, Y^{(m)}|X^{(o)}\right] dY^{(m)} \\ &= \left[R|Y^{(o)}, X^{(o)}\right] \int \left[Y^{(o)}, Y^{(m)}|X^{(o)}\right] dY^{(m)} \\ &= \left[R|Y^{(o)}, X^{(o)}\right] \left[Y^{(o)}|X^{(o)}\right], \end{aligned} \quad (6)$$

where we obtain (6) under an MAR assumption given $X^{(o)}$,

$$Pr\left(R|Y^{(o)}, Y^{(m)}, X\right) = Pr\left(R|Y^{(o)}, X^{(o)}\right).$$

If we assume that $[R|Y^{(o)}, X^{(o)}]$ is free of β then we can consistently estimate β using just $[Y^{(o)}|X^{(o)}]$ and maximum likelihood. To do so we need to specify $[Y^{(o)}|X^{(o)}]$, but this is a problem because we construct models for Y (and hence $Y^{(o)}$) based on X.

To assess the effect of conditioning on just the observed covariates $X^{(o)}$, we decompose $[Y^{(o)}|X^{(o)}]$ as

$$\begin{aligned}\left[Y^{(o)}|X^{(o)}\right] &= \int \left[Y^{(o)}, X^{(m)}|X^{(o)}\right] dX^{(m)} \\ &= \int \left[Y^{(o)}|X^{(m)}, X^{(o)}\right] \left[X^{(m)}|X^{(o)}\right] dX^{(m)} \\ &= \int \left[Y^{(o)}|X\right] \left[X^{(m)}|X^{(o)}\right] dX^{(m)}. \end{aligned} \quad (7)$$

Equation (7) shows that $[Y^{(o)}|X^{(o)}]$ is $[Y^{(o)}|X]$ mixed by $[X^{(m)}|X^{(o)}]$. It is easy to construct examples where $[Y^{(o)}|X^{(o)}] \neq [Y^{(o)}|X)]$. Thus, we cannot immediately use models for $[Y^{(o)}|X]$ as models for $[Y^{(o)}|X^{(o)}]$. In particular, (7) shows that to model $[Y^{(o)}|X^{(o)}]$ we need to model the missingness process to obtain $[X^{(m)}|X^{(o)}]$.

2.3. Consistent estimation based on $X^{(o)}$ under MAR

We can obtain consistent estimates of β based only on $X^{(o)}$ with observations MAR in several settings. In particular, if the components of $X^{(o)}$ are exactly the same as we would

have calculated with no missing data then the conditioning sets in $[Y^{(o)}|X]$ and $[Y^{(o)}|X^{(o)}]$ will be the same and models for $[Y^{(o)}|X]$ immediately apply to observed data. Therefore, we will obtain consistent estimation of β using $\hat{\beta}_W$ in this setting.

There are several situations where X has missing values but the between/within cluster decomposition is the same as with complete X. First, $\bar{X}_i$ may be known despite missing X_{ij}. Examples include cross-over studies and studies with planned visit times. Also, large national surveys may have known cluster-averages of quantities such as neighborhood socio-economic status but have missing individual values. Another important situation is one where we base covariate decompositions on X_{i1} and $X_{ij} - X_{i1}$, i.e., baseline and change over baseline. It is not unusual for longitudinal studies to require complete baseline data in order to enter a subject into the study.

Consider decomposing a covariate vector with no missing values $(X_{i1},\ldots,X_{iT_i})$ into between- and within-subject components based on X_{i1}. This yields the vector

$$X_i^1 = \left(X_{i1}, 0, X_{i2} - X_{i1}, \ldots, X_{iT_i} - X_{i1}\right).$$

With missing values, the between- and within-subject components based on the observed data are exactly the same as we would have calculated with no missing data. This is not the case when we decompose based on the observed mean of the covariates $\bar{X}_i^{(o)}$. With covariate decompositions based on observed $\bar{X}_i^{(o)}$, the components of $X^{(o)}$ are not the components of complete X. The complete covariates are

$$X_i^t = \left(\bar{X}_i^t, X_{i1} - \bar{X}_i^t, \ldots, X_{iT_i} - \bar{X}_i^t\right),$$

where the superscript t denotes total and $\bar{X}_i^t = \sum_{j=1}^{T_i} X_{ij}/T_i$, while the observed covariates are

$$X_i^{(o)} = \left(\bar{X}_i^{(o)}, X_{i1} - \bar{X}_i^{(o)}, \ldots, X_{in_i} - \bar{X}_i^{(o)}\right).$$

It is well-known that $E\left[\bar{X}_i^{(o)}\right] \neq E\left[\bar{X}_i^t\right]$ when X has observations MAR. This leads to systematic differences between $\bar{X}_i^{(o)}$ and $\bar{X}_i^t$, and thus to potentially biased estimates.

To summarize, when $cor(X_{ij}, b_i) \neq 0$ and we have data MAR, we can obtain consistent estimation of β with complete covariate information, X, using between/within subject covariate decomposition methods. We can also obtain consistent estimation of β in cases

where decomposition based on observed data only coincides with that using the complete covariate information. In other cases, as we show by simulation in the next section, the between/within covariate decomposition methods using only the observed covariates behave like conditional maximum likelihood and can be inconsistent.

3. Simulation studies

3.1. Structure

We carried out simulation studies patterned after the SALSA study to evaluate the performance of mixed-effects model methods, including those with covariates decomposed into between- and within-subject components, in settings with observations missing at random and with $cor(X_{ij}, b_i) \neq 0$. We generated data from linear mixed-effects, mixed-effects logistic and mixed-effects Poisson models involving the predictors age, group and age by group interaction to allow the estimation of covariate effects commonly of interest in longitudinal studies. Specifically, we generated longitudinal responses with random intercepts and slopes from models with linear predictor

$$\eta_{ij}=\beta_0+b_{0i}+(\beta_{AGE}+b_{1i})\,AGE_{ij}+\beta_T TRT_i+\beta_I TRT_i \times AGE_{ij}, \quad (8)$$

$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \overset{ind}{\sim} N\left(\begin{bmatrix} \delta AGE_{i1} \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{b0} & \sigma_{12} \\ \sigma_{12} & \sigma^2_{b1} \end{bmatrix}\right),$$

where i=1,…, 500, j=1,…, 7, to correspond to 7 annual visits, TRT_i is a binary group variable equal to 1 for one half of the subjects and 0 for the others, AGE_{i1}~N(0, 1) and AGE_{ij} increases by 1/7 at each visit. The mixed-effects models (8) included random intercepts, b_{0i}, and slopes, b_{1i}, to produce subject-to-subject variability in the initial response (e.g. cognitive function) and change in the response over time, respectively. The parameter δ produces association between the random intercept, b_{0i} and the covariate AGE_{ij}. Each simulation generated 2000 data sets and the true parameter values varied by outcome type. The true parameter values for linear mixed-effects models were β_0=140, β_{AGE}=−1, β_{TRT} =5.0, β_I =1.0, δ=−0.5, σ_{b0} =2.0, σ_{b1} =1.0, σ_{12}=−1.0. The true parameter values for mixed-effects logistic models were β_0=−3.0, β_{AGE} =0.2, β_{TRT} =0.3, β_I =0.2, δ=1.2, σ_{b0} =√6, σ_{b1} =1.0, σ_{12}=−0.5√6. The true parameter values for Poisson mixed-effects models were β_0=−1.0, β_{AGE} =0.2, β_{TRT} =0.25, β_I =0.3, δ=1.2, σ_{b0}=1.0, σ_{b1}=0.25, σ_{12}=−0.125.

After generating a data set and for j > 1, we generated `missing value' indicators R_{i2}, …,R_{i7} using a logistic model

$$logit\,\{Pr\,(R_{ij}=1)\}=\gamma_0+\gamma_y y^{\mathscr{S}}_{ij-1}+\gamma_A AGE_{ij-1}+\gamma_T TRT_i+\gamma_I TRT_i \times AGE_{ij-1}. \quad (9)$$

The superscript $\mathscr{S}$ denotes standardized variates with mean=0 and standard deviation=1. In the simulations, we always observed Y_{i1}, as well as the predictors AGE_{i1} and TRT_i. After the first generated R_{ij}=0 for a subject, we set all subsequent R_{ik} =0, k>j to generate a monotone drop-out process. The true values of the missingness model (9) parameters for the linear mixed-effects model simulations were: γ_0=2.0, γ_Y =0.5, γ_A=−0.5, γ_T =0.5, γ_I =−0.5. The true

values of the missingness model (9) parameters for the mixed-effects logistic model simulations were: $\gamma_0 = 2.0$, $\gamma_Y = 0.5$, $\gamma_A = -0.5$, $\gamma_T = 0.5$, $\gamma_I = -0.5$. The true values of the missingness model (9) parameters for the mixed-effects Poisson model simulations were: $\gamma_0 = 2.0$, $\gamma_Y = -0.7$, $\gamma_A = -0.7$, $\gamma_T = 0.5$, $\gamma_I = 0.5$.

3.2. Fitting methods

We fitted five different models/approaches to each generated data set:

1. Standard maximum likelihood (ML) with linear predictor

$$\eta_{ij} = \beta_0 + b_{0i} + (\beta_{AGE} + b_{1i})\, AGE_{ij} + \beta_G TRT_i + \beta_I TRT_i \times AGE_{ij}. \quad (10)$$

2. Between/within using the true average for subject i, $\overline{AGE}_i^t$, i.e. based on all ages generated before the missing data process. Explicitly, we replaced $(\beta_{Age} + b_{1i})AGE_{ij}$ in η_{ij} (10) by

$$\left(\beta_{BAge} + b_{1i}\right) \overline{AGE}_i^t + \left(\beta_{WAge} + b_{1i}\right) \left(AGE_{ij} - \overline{AGE}_i^t\right)$$

and replaced $\beta_I TRT_i \times AGE_{ij}$ by

$$\beta_{IB} TRT_i \times \overline{AGE}_i^t + \beta_{IW} TRT_i \times \left(AGE_{ij} - \overline{AGE}_i^t\right).$$

3. Between/within using AGE_{i1}, i.e. based on baseline ages. Explicitly, we replaced $(\beta_{Age} + b_{1i})\, AGE_{ij}$ in η_{ij} (10) by

$$\left(\beta_{BAge} + b_{1i}\right) AGE_{i1} + \left(\beta_{WAge} + b_{1i}\right) (AGE_{ij} - AGE_{i1})$$

and replaced $\beta_I TRT_i \times AGE_{ij}$ by

$$\beta_{IB} TRT_i \times AGE_{i1} + \beta_{IW} TRT_i \times (AGE_{ij} - AGE_{i1}).$$

4. Between/within using the observed average for subject i, $\overline{AGE}_i^{(o)}$, i.e. based only on observed ages after missing data process. Explicitly, we replaced $(\beta_{Age} + b_{1i})AGE_{ij}$ in η_{ij} (10) by

$$\left(\beta_{BAge} + b_{1i}\right) \overline{AGE}_i^{(o)} + \left(\beta_{WAge} + b_{1i}\right) \left(AGE_{ij} - \overline{AGE}_i^{(o)}\right)$$

and replaced $\beta_I TRT_i \times AGE_{ij}$ by

$$\beta_{IB} TRT_i \times \overline{AGE}_i^{(o)} + \beta_{IW} TRT_i \times \left(AGE_{ij} - \overline{AGE}_i^{(o)}\right).$$

5. Standard conditional likelihood approaches (McCullagh & Nelder 1989) obtained by computing likelihoods conditional on sufficient statistics for random intercepts.

We fitted linear mixed-effects models using Proc Mixed in SAS (SAS Inc., Cary, NC, USA) and mixed-effects logistic and mixed-effects Poisson models using Proc Nlmixed in SAS, making the assumption that the random intercepts and slopes followed a bivariate Normal distribution. We fitted conditional likelihood approaches for Normal and Poisson responses using a fixed effects model approach (Allison 2005) that added fixed, subject-specific intercepts to standard linear and Poisson regression models. To implement the Poisson conditional likelihood approach, we also dropped subjects whose responses were all zeros and subjects who provided only a single (baseline) observation. We fitted the conditional likelihood approach for logistic models using Proc Logistic in SAS.

3.3. Results

Tables 2, 3 and 4 present means and standard deviations of parameter estimates for linear, logistic and Poisson mixed-effects models, respectively. As Neuhaus & McCulloch (2006) showed, the estimates from standard linear mixed-effects models in Table 2 of covariate effects, such as *AGE* and the *AGE* by *TRT* interaction effects, that are correlated with the random intercepts b_{i0}, are biased due to the correlation between b_{i0} and baseline age, AGE_{i1}. In line with the theory presented in Section 2, which shows consistency of the within parameter estimators, the within-subject components of the decomposition methods based on the true average for subject *i*, $\overline{AGE}_i^t$ and baseline age, AGE_{i1} produce essentially no bias in the estimates of the main and interaction effects of age. The between-subject components of the decomposition methods differ substantially from the true values again due to the correlation between the random intercepts b_{i0} and baseline age AGE_{i1}. As expected from the theory presented in Section 2, decomposition methods based on the true covariate average and the baseline covariate produced nearly identical estimates and associated standard deviations. As suggested by the theory from Section 2, and consistent with previous work (Rathouz 2004; Roy *et al*. 2006) the within-subject estimators produced by the decomposition methods based on the observed average for subject *i*, $\overline{AGE}_i^{(o)}$ and the conditional likelihood estimators are biased for the true values because observations are missing at random (MAR). It is also noteworthy that the decomposition methods based on the observed average and the conditional likelihood approach produced nearly identical estimates and associated standard deviations.

The results in Tables 3 and 4 for mixed-effects logistic and Poisson models, respectively, follow the same pattern as those in Table 2. Standard mixed-effects models produce biased estimates due to the correlation between the random effects and covariates. The within-subject components of the decomposition methods based on the true average for subject *i*, $\overline{AGE}_i^t$ and baseline age, AGE_{i1} produce essentially no bias (and are similar to one another) in the estimates of the main and interaction effects of age. The within-subject components of the decomposition methods based on the observed average for subject *i*, $\overline{AGE}_i^{(o)}$. The conditional likelihood estimators are biased (and are similar to one another) for the true values because observations are missing at random (MAR). As in Table 2, the

decomposition methods based on the true covariate average and the baseline covariate produced nearly identical estimates and associated standard deviations while decomposition methods based on the observed average and the conditional likelihood approach also produced nearly identical estimates.

4. Example: poor cognitive functioning, physical functioning and age

We further illustrate our results by fitting several mixed-effects logistic model approaches to data from the longitudinal SALSA study. The outcome of interest is a binary indicator of poor cognitive functioning, 3MSE<80. The predictors of interest are activities of daily living (ADL), a count of the number of activities such as ability to feed, bathe, dress, and groom oneself that the subject could not perform, and age. As we noted in Section 1, SALSA measured outcomes and predictors at up to seven visits over a 10 year period and missing data, as well as subject drop-out, were common.

We fitted four mixed-effects logistic model approaches to these data:

1. Standard mixed-effects logistic model with predictors ADL and age;
2. Between/within covariate decomposition for each predictor using the observed average;
3. Between/within covariate decomposition for each predictor using the first value; and
4. Standard conditional likelihood.

Table 1 presents the estimated covariate effects, along with associated standard errors for each of the fitting approaches. A likelihood ratio test rejects the null hypothesis of common between- and within-subject effects of ADL and age, $H_0 : \beta_{ADL,W} = \beta_{ADL,B}$, $\beta_{ADL,W} = \beta_{ADL,B}$, and a subsequent post-hoc test indicates that $\beta_{ADL,W} \neq \beta_{ADL,B}$, i.e. the within-subject increase in Pr(poor cognitive function) as a person becomes more disabled, differed from differences in Pr(poor cognitive function) between persons who started the study with different physical function. A post-hoc test using the decomposition of age based on the observed average indicates that $\beta_{ADL,W} \neq \beta_{ADL,B}$, while the difference between the between- and within-subject effects of age based on the first value are not statistically significant. Differences between $\beta_{ADL,W}$ and $\beta_{ADL,B}$ indicate that the magnitude of the within-subject rate of change over time in poor cognitive functioning differs from the magnitude of the difference in rates of poor cognitive functioning between subjects who differ in age. The observed differences between $\hat{\beta}_{ADL}$ and $\hat{\beta}_{ADL,W}$ are expected from the correlation of ADL measurements and predicted underlying cognitive functioning observed in Figure 3. The correlation in Figure 3 may be due to omitted covariates that are associated with cognitive functioning and ADL measurements and may be evidence of confounding. The discrepancies between $\hat{\beta}_{ADL,W}$ and $\hat{\beta}_{ADL,B}$, as well as those between $\hat{\beta}_{AGE,W}$ and $\hat{\beta}_{AGE,B}$, lead to differences between $\hat{\beta}_{ADL,W}$ and $\hat{\beta}_{ADL}$, as well as to differences between $\hat{\beta}_{AGE,W}$ and $\hat{\beta}_{AGE}$, and suggest that both $\hat{\beta}_{ADL}$ and $\hat{\beta}_{AGE}$ are biased. Consistent with the theory in Section 2 and simulations in Section 3, the

conditional maximum likelihood estimates and the within-subject estimates $\hat{\beta}_{ADL,W}$ and $\hat{\beta}_{AGE,W}$ based on covariate decomposition methods using observed means $\overline{ADL}^{(o)}$ and $\overline{AGE}^{(o)}$, respectively, were nearly identical but somewhat different from the within-subject estimates based on ADL_{i1} and AGE_{i1}.

5. Discussion

This paper presents an easy-to-use approach to obtain consistent estimates of change, in settings with cluster-level confounding and observations missing at random. One can fit the approach using standard mixed model software and use any generalized linear model of interest. Essentially, our approach provides a simple approach to obtaining consistent conditional likelihood-like estimates for a wide variety of link functions in settings with observations missing at random. In contrast, the approaches of Rathouz (2004) and Roy *et al.* (2006) require specialized software and apply to only logistic or Poisson models, respectively. In addition to being simple to implement, decomposing covariates using baseline values also provides scientifically appealing covariate effect estimates; scientific interest often focuses on the magnitude of change in an outcome from baseline and the association of these changes with changes from baseline in predictors.

Acknowledgement

Grants CA082370 and AG12975 from the U.S. National Institutes of Health supported this research. We thank Dr. Mary Haan of the University of California, San Francisco for providing the data from the Sacramento Area Latino Study on Aging (SALSA) and Dr. Ross Boylan of the University of California, San Francisco for assistance with computing.

References

Allison, P. Fixed Effects Regression Methods for Longitudinal Data Using SAS. SAS Institute; Cary, NC: 2005.

Enders, CK. Centering predictors and contextual effects. In: Scott, MA.; Simonoff, JS.; Marx, BD., editors. The SAGE handbook of multilevel modeling. SAGE Publications; London: 2013. p. 88-107.

Haan M, Mungas D, Gonzalez H, Ortiz T, Acharya A, Jagust W. Prevalence of dementia in older latinos: the influence of type 2 diabetes mellitus, stroke and genetic factors. J. Amer. Geriatr. Soc. 2003; 51:169–177. [PubMed: 12558712]

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

McCullagh, P.; Nelder, J. Generalized Linear Models. Chapman and Hall; London: 1989.

McCulloch, CE.; Searle, SR.; Neuhaus, JM. Generalized, Linear and Mixed Models. 2nd edn. Wiley; New York: 2008.

Neuhaus JM, Kalbfleisch JD. Between- and within-cluster covariate effects in the analysis of clustered data. Biometrics. 1998; 54:638–645. [PubMed: 9629647]

Neuhaus JM, McCulloch CE. Separating between- and within-cluster covariate effects using conditional and partitioning methods. J. R. Statist. Soc. Ser. B Statist. Methodol. 2006; 68:859–872.

Rathouz P. Fixed effect models for longitudinal binary data with dropouts missing at random. Statist. Sinica. 2004; 14:969–988.

Roy J, Alderson D, Hogan JW, Tashima KT. Conditional inference methods for incomplete Poisson data with endogenous time-varying covariates: Emergency department use among HIV-infected women. J. Amer. Statist. Assoc. 2006; 101:434–434.

Skrondal A, Rabe-Hesketh S. Protective estimation of mixed-effects logistic regression when data are not missing at random. Biometrika. 2014; 101:175–188.

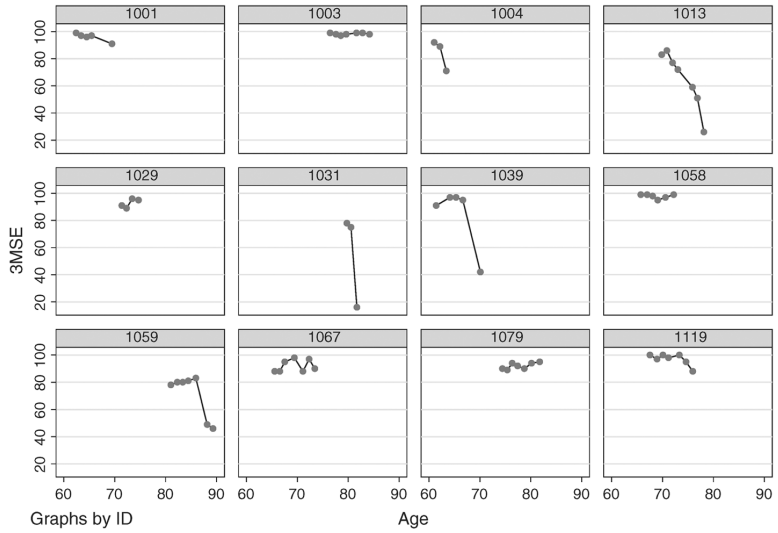

Figure 1.

Subject-specific trajectories in Modified Mini-Mental State Examination (3MSE) for the Sacramento Area Latino Study on Aging (SALSA) study.

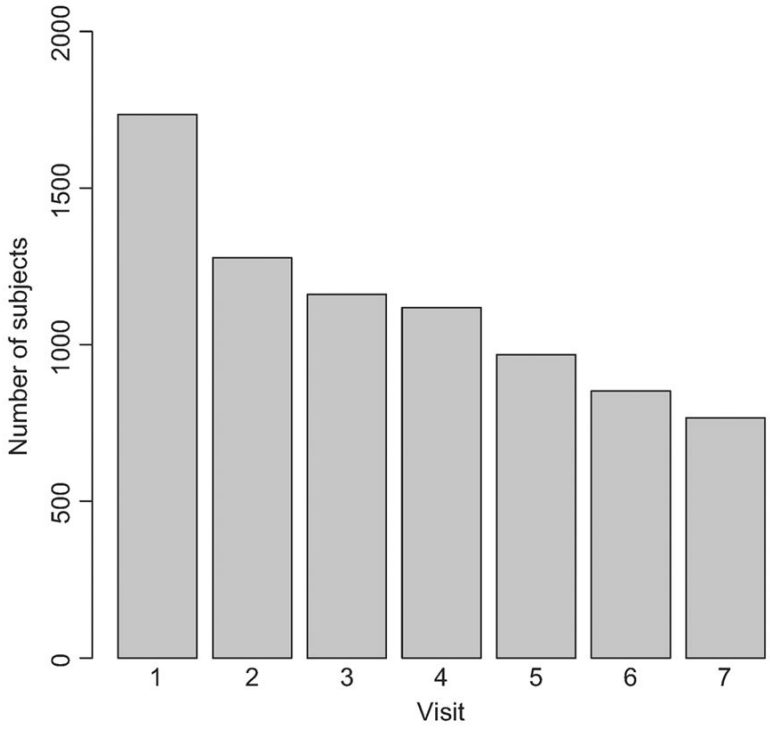

Figure 2.
Number of subjects contributing data at each visit in the Sacramento Area Latino Study on Aging (SALSA).

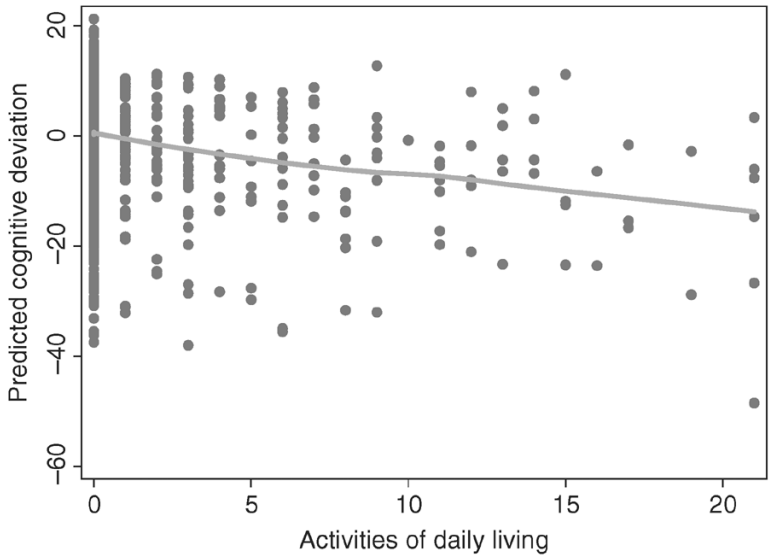

Figure 3.

Association of random intercepts with Activities of Daily Living (ADL) at time 1. Solid line is a locally weighted scatterplot smooth (LOWESS).

Table 1

Activities of Daily Living (ADL) and age effect estimates from four methods fitted to the Sacramento Area Latino Study on Aging (SALSA) data. Standard errors in parentheses.

	$\hat{\beta}_{ADL}$	$\hat{\beta}_{ADL,W}$	$\hat{\beta}_{ADL,B}$
Standard ML	0.13 (0.02)		
Between/within, ADL_{i1}		0.08 (0.02)	0.33 (0.04)
Between/within, observed $\overline{ADL}_i^{(o)}$		0.05 (0.02)	0.29 (0.03)
Conditional likelihood		0.05 (0.02)	

	$\hat{\beta}_{Age}$	$\hat{\beta}_{Age,W}$	$\hat{\beta}_{Age,B}$
Standard ML	0.16 (0.01)		
Between/within, AGE_{i1}		0.17 (0.02)	0.14 (0.01)
Between/within, observed $\overline{AGE}_i^{(o)}$		0.22 (0.02)	0.11 (0.02)
Conditional likelihood		0.23(0.02)	

Table 2

Observed means and standard deviations, in parentheses, of the regression coefficients of several methods for fitting linear mixed models with random intercepts and slopes to simulated longitudinal data with observations missing at random and correlation between the random intercepts and the predictor AGE at time 0.

Approach	β_{AGE}	β_{AGEW}	β_{AGEB}	β_{TRT}	β_{I}	β_{IW}	β_{IB}
True	−1.0			5.0	1.0		
Standard LME	−5.39 (0.97)			5.09 (2.01)	1.47 (1.31)		
Bet/With $\overline{AGE}_i^t$		−1.00 (1.20)	−11.00 (1.34)	5.07 (2.11)		1.00 (1.61)	0.95 (1.90)
Bet/With AGE_{i1}		−1.00 (1.20)	−11.00 (1.34)	5.05 (1.93)		1.00 (1.61)	0.95 (1.90)
Bet/With $\overline{AGE}_i^{(o)}$		−1.73 (1.22)	−10.52 (1.45)	5.50 (2.04)		1.26 (1.63)	0.51 (2.08)
Cond like		−1.81 (1.21)				1.28 (1.63)	

The fitting approaches were: (i) Standard linear mixed-effects models that ignored associations between random intercepts and predictors; (ii) Between/within using the true average for subject *i*, $\overline{AGE}_i^t$; (iii) Between/within using AGE_{i1}; (iv) Between/within using the observed average for subject *i*, $\overline{AGE}_i^{(o)}$; (v) Conditional likelihood.

Table 3

Observed means and standard deviations, in parentheses, of the regression coefficients of several methods for fitting mixed-effects logistic models with random intercepts and slopes to simulated longitudinal data with observations missing at random and correlation between the random intercepts and the predictor AGE at time 0.

Approach	β_{AGE}	β_{AGEW}	β_{AGEB}	β_{TRT}	β_{I}	β_{IW}	β_{IB}
True	0.2			0.3	0.2		
Std Mixed Logistic	2.57 (0.34)			0.20 (0.38)	0.19 (0.37)		
Bet/With $\overline{AGE}_i^t$		0.18 (0.48)	3.16 (0.37)	0.31 (0.45)		0.23 (0.61)	0.21 (0.42)
Bet/With AGE_{i1}		0.18 (0.48)	3.17 (0.37)	0.30 (0.40)		0.23 (0.61)	0.31 (0.43)
Bet/With $\overline{AGE}_i^{(o)}$		−0.27 (0.49)	3.30 (0.37)	0.20 (0.43)		0.31 (0.63)	0.31 (0.43)
Cond Like		−0.51 (0.47)				0.30 (0.63)	

The fitting approaches were: (i) Standard mixed-effects logistic models that ignored associations between random intercepts and predictors; (ii) Between/within using the true average for subject *i*, $\overline{AGE}_i^t$; (iii) Between/within using AGE_{i1}; (iv) Between/within using the observed average for subject *i*, $\overline{AGE}_i^{(o)}$; (v) Conditional likelihood.

Table 4

Observed means and standard deviations, in parentheses, of the regression coefficients of several methods for fitting mixed-effects Poisson models with random intercepts and slopes to simulated longitudinal data with observations missing at random and correlation between the random intercepts and the predictor AGE at time 0.

Approach	β_{AGE}	β_{AGEW}	β_{AGEB}	β_{TRT}	β_{I}	β_{IW}	β_{IB}
True	0.2			0.25	0.3		
Std Mixed Poisson	0.99 (0.10)			0.33 (0.16)	0.17 (0.15)		
Bet/With $\overline{AGE}_i^t$		0.20 (0.13)	1.40 (0.10)	0.23 (0.17)		0.31 (0.16)	0.33 (0.15)
Bet/With AGE_{i1}		0.20 (0.13)	1.40 (0.10)	0.24 (0.14)		0.31 (0.16)	0.33 (0.15)
Bet/With $\overline{AGE}_i^{(o)}$		0.35 (0.13)	1.32 (0.11)	0.21 (0.16)		0.26 (0.16)	0.34 (0.17)
Cond Like		0.35 (0.14)				0.26 (0.17)	

The fitting approaches were: (i) Standard mixed-effects Poisson models that ignored associations between random intercepts and predictors; (ii) Between/within using the true average for subject i, $\overline{AGE}_i^t$; (iii) Between/within using AGE_{i1} (iv) Between/within using the observed average for subject i, $\overline{AGE}_i^{(o)}$; (v) Conditional likelihood.