Statistica Sinica **14**(2004), 969-988

FIXED EFFECTS MODELS FOR LONGITUDINAL BINARY DATA WITH DROP-OUTS MISSING AT RANDOM

Paul J. Rathouz

University of Chicago

Abstract: We consider the problem of attrition under a logistic regression model for longitudinal binary data in which each subject has his own intercept parameter, and where parameters are eliminated via conditional logistic regression. This is a fixed-effects, subject-specific model which exploits the longitudinal data by allowing subjects to act as their own controls. By modeling and conditioning on the drop-out process, we develop a valid but inefficient conditional likelihood using the complete-record data. Then, noting that the drop-out process is ancillary in this model, we use a projection argument to develop a score with improved efficiency over the conditional likelihood score, and embed both of these scores in a more general class of estimating functions. We then propose a member of this class that approximates the projected score, while being much more computationally feasible. We study the efficiency gains that are possible using a small simulation, and present an example analysis from aging research.

Key words and phrases: Attrition, conditional likelihood, Conditional logistic regression, missing data, nuisance parameter, projection, subject-specific model.

1. Introduction

An important strength of longitudinal data is that subjects can act as their own controls in evaluating the effects of treatments, policy interventions, and other time-varying exposures on outcomes. Longitudinal study designs eliminate confounding that can arise in cross-sectional studies between such exposures and other subject level factors (Diggle, Liang and Zeger (1994, Chap.1)).

For longitudinal binary outcome data, a common model is

$$\text{logit}\{\text{E}\,(Y_{it}|\mathbf{X}_i)\} = q_i + X'_{it}\beta, \tag{1}$$

where Y_{it} is a binary outcome variable on subject i at time t, X_{it} is a vector of time-varying covariates, β is a vector of regression coefficients, and $\mathbf{X}_i$ is the matrix $\mathbf{X}_i = (X_{i1}, \ldots, X_{iJ})'$. In this logistic model, q_i is a subject-specific intercept that accounts for the fact that the components of the vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})'$ are repeated measures on a single subject i, and hence are positively correlated. Models of this form are sometimes referred to as 'subject-specific models' (Zeger,

Liang and Albert (1988)) because the interpretation of β is the effect of X_{it} on Y_{it}, adjusting for all factors figuring into the subject-specific intercept q_i.

These models are sometimes fitted by assuming that the q_i's are random quantities that follow a probability law such as the normal distribution (e.g., Breslow and Clayton (1993)). An alternative is to assume that the q_i's are fixed unknown parameters (Greene (2003) and Diggle, Liang and Zeger (1994, Section 9.3.1)). While both models yield a subject-specific interpretation of β, the 'random effects' and 'fixed effects' approaches differ fundamentally. In the random effects models, it is generally assumed that q_i is independent of the matrix of covariates $\mathbf{X}_i$, although it is certainly possible to model the dependence of q_i on $\mathbf{X}_i$. By contrast, in fixed effects models, q_i captures all subject level factors, including those related to $\mathbf{X}_i$. This yields inferences about the effects of X_{it} on Y_{it} that are automatically adjusted for confounding due to subject-level factors. In this way, subjects act as their own controls in longitudinal studies. Greene (2003, p.700) illustrates the role of fixed effects models in the analysis of longitudinal binary data through an analysis of time-varying factors on product innovation in a longitudinal sample of 1270 German firms (Bertschek and Lechner (1998)).

Typically, when q_i is seen as a fixed quantity, model (1) is estimated via conditional logistic regression (CLR; Breslow and Day (1980)). CLR eliminates q_i from the ith subject's likelihood contribution by conditioning on the sum $\sum_t Y_{it}$, which is a sufficient statistic for q_i. A third approach midway between the random and fixed effects models is to assume that q_i is a random variable with an unknown distribution that depends on $\mathbf{X}_i$ in an unknown way. This yields a semiparametric model, with $X_{it}'\beta$ being the parametric part, and the distribution of q_i given $\mathbf{X}_i$ being the nonparametric. The CLR estimator is semiparametric efficient for β in this model (Rathouz (2003)).

To fix ideas, we consider an example from aging research in which interest is on the question of whether elderly subjects with weaker memory experience greater subsequent increases in disability The data used to address this question are from the Study of Assets and Health Dynamics Among the Oldest Old (AHEAD; Soldo, Hurd, Rodgers and Wallace (1997)), a U.S. national sample of subjects 70 years and older, and were collected over four waves. They include a baseline measure of memory and a longitudinal binary measure of disability. A concern in a cross-sectional analysis of this problem is that unobserved confounding factors may lead to a spurious association between memory and disability. The question is more appropriately addressed with a fixed effects model with the longitudinal disability measure as the response, and the interaction of time (year) with baseline memory as the key predictor of interest. An analysis of these data is presented in Section 5.

Attrition is an important problem in aging research such as this. For example, our data set comprises 6,350 subjects, but only 3,705 are assessed at all four time points, and 887 dropped out of the study after the first wave. Subjects may drop out because they become too disabled to participate, because they move into an assisted living or nursing facility, or because they die; often, the investigators do not know which if any of these events has occurred for a subject who can not be located at follow-up waves. Furthermore, the reason for dropout is potentially related to the longitudinal outcome of disability. Bias due to attrition is therefore a serious concern.

As this example suggests, problems can arise when subjects drop out of the study after only T_i observations, where, for some subjects, $T_i < J$. When the drop-out time T_i is independent of both $\mathbf{Y}_i$ and $\mathbf{X}_i$, analysis based on the standard conditional likelihood generated from the first T_i observations will yield consistent and asymptotically normal estimators for β. However, if drop-out depends on $\mathbf{Y}_i$ and/or $\mathbf{X}_i$, a standard complete record analysis based on the first T_i observations can yield biased estimators. In this paper, we consider drop-outs that are missing at random, where T_i may depend on past data $Y_{i1}, \ldots, Y_{iT_i}$, but not on future data $Y_{i,T_i+1}, \ldots, Y_{iJ}$. Specifically, we assume that

$$I(T_i = t) \amalg Y_{i,t+1}, \ldots, Y_{iJ} | Y_{i1}, \ldots, Y_{it}, \mathbf{X}_i, Z_i, \tag{2}$$

where $I(\cdot)$ is the indicator function. We refer to this condition as 'missing at random drop-out' (Little (1995)). In (2), Z_i is a vector of subject-level variables which may effect drop-out time. Condition (2) arises naturally under a hazard model for drop-out which expresses the probability that $T_i = t$ given $T_i \geq t$ as a function of $Y_{i1}, \ldots, Y_{it}, \mathbf{X}_i, Z_i$.

An alternative identifying assumption to (2) is the slightly more general condition

$$I(T_i = t) \amalg Y_{i,t+1}, \ldots, Y_{iJ}, X_{i,t+1}, \ldots, X_{iJ} | Y_{i1}, \ldots, Y_{it}, X_{i1}, \ldots, X_{it}, Z_i,$$

which does not require the X_{it}'s to be measured after drop-out time T_i, and which is therefore applicable to problems with random time-varying covariates. The methods presented in this paper extend easily to this case, but the notation is more cumbersome; a sketch of the development is given in Section 6.

Models are generally identifiable under missing at random (MAR) assumptions such as (2) (Little and Rubin (1987)). However, a full likelihood approach to the fixed effects modeling problem considered here would require integrating over the missing $Y_{i,T_i+1}, \ldots, Y_{iJ}$, and, by (1), the distribution of these missing data elements depends on the unknown q_i. As this approach does not appear feasible, we develop a method wherein we condition on T_i and base inferences

on the conditional likelihood obtained by further conditioning on $\sum_{t=1}^{T_i} Y_{it}$. This approach exploits a model for the drop-out process to compute the likelihood conditional on T_i. Development of this bias-corrected 'complete-record' conditional likelihood is the subject of Section 2.

Other authors have developed methods to handle attrition in longitudinal studies; Little (1995) provides a review. Most of these are based either on 'marginal models', wherein primary interest is on the mean of Y_{it} as a function of X_{it} (Robins, Rotnitzky and Zhao (1995), Robins and Rotnitzky (1995), Fitzmaurice, Molenberghs and Lipsitz (1995), Baker (1995) and Diggle and Kenward (1994)), or on random effects models (e.g., Wu and Carroll (1988) and Ten Have, Kunselman, Pulkstenis and Landis (1998)). Attrition in fixed effects models has received less attention. Conaway (1992) considers a general class of polytomous data fixed effects models with missing responses, of which our model is a special case. However, his method of handling attrition differs from ours in that it is based on application of the EM algorithm to the full-data conditional likelihood. By contrast, the method we present in Section 2 is based on a conditional likelihood constructed using only complete-records.

A problem with the complete-record approach is that it may result in inefficient inferences for β, primarily because it relies on information in the observed drop-out process, which is ancillary for β. In Section 3, we address this problem by identifying a class of estimating functions of which the score function developed in Section 2 is one member. Using a projection argument, we identify the efficient member of that class which, in particular, is guaranteed to improve efficiency in β-estimation over the score in Section 2. Finally, as the efficient estimating function is difficult to compute, we propose an approximation to it that simplifies computation considerably. Section 4 contains a small simulation study, and in Section 5 we illustrate our methods with an analysis of the AHEAD data discussed above.

The method developed here follows a similar program to that in Rathouz, Satten and Carroll (2002). In that paper, we elaborated a methodology for handling missing covariates in conditional logistic regression models. This paper differs in that it considers longitudinal data, which specifically incorporates the element of time, and focuses on missing responses rather than missing covariates. As such, the bias-corrected estimator proposed in Section 2 takes a different form than that in our previous work. Moreover, the projection argument and subsequent approximation in Sections 3.3−3.4 are completely new and specifically developed for longitudinal data.

2. Complete-Record Analysis of Fixed Effects Models

2.1. Data, notation and model of interest

Consider a random sample of subjects $i = 1, \ldots, K$, and suppose that each subject is potentially assessed at J times, denoted by $t = 1, \ldots, J$. Assessment t on subject i yields data (Y_{it}, X_{it}), where X_{it} is a vector of time-varying covariates and Y_{it} is a binary response variable. For ease of exposition, we assume that each subject has the same set of equally-spaced potential assessment times $t = 1, \ldots, J$, although the methods to be developed would certainly allow for different sets of such times across subjects. In addition to Y_{it} and X_{it}, let Z_i denote a vector of time-constant subject-level covariates. While these covariates will not figure directly into our model of interest, they are included in the data structure because they may figure into the generation of time-varying covariates X_{it} or into the drop-out process. For example, time-varying covariates X_{it} may be deterministic functions of time, such as t or $Z_i \times t$. Alternatively, X_{it} may include measures of exogenous time-varying processes such as the level of ambient air pollution in subject i's ZIP code at time t. For ease of exposition, we operate under (2) and assume that X_{it} can be measured even if subject i drops out before t. Extension to the case where X_{it} is a random time-varying covariate that cannot be measured after drop-out is conceptually straightforward; details are given in the discussion.

To account for random drop-out across subjects, assume that subject i is observed only at times $t = 1, \ldots, T_i \leq J$, where $T_i \geq 1$ for all i. We refer to T_i as the ith subject's 'drop-out time'. Let $R_{it} = 1$ or 0 indicate whether the tth observation (Y_{it}, X_{it}) is observed or missing for the ith subject. That is, that $R_{it} = 1$ iff $t \leq T_i$.

To denote the data for a given subject i, write $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})'$ for the $J \times 1$ vector of binary responses, including any values missing due to drop-out. Similarly define the $J \times p$ matrix $\mathbf{X}_i = (X_{i1}, \ldots, X_{iJ})'$ of $p \times 1$ row vectors of covariates for subject i, and the vector $\mathbf{R}_i$ of missing data indicators. Further define $\mathbf{Y}_{i,\text{obs}}$ to be the T_i observed components of $\mathbf{Y}_i$, and $\mathbf{X}_{i,\text{obs}}$ to be the T_i observed rows of $\mathbf{X}_i$, before or at drop-out. Note that which components of $\mathbf{Y}_i$ and $\mathbf{X}_i$ appear in $\mathbf{Y}_{i,\text{obs}}$ and $\mathbf{X}_{i,\text{obs}}$ is a function of the indicator vector $\mathbf{R}_i$. Finally, a subscript t added to a $\mathbf{Y}_i$ or $\mathbf{X}_i$ is used to denote the sub-vector or sub-matrix comprising the first t components of a vector or first t components of a matrix. For example, $\mathbf{Y}_{it} = (Y_{i1}, \ldots, Y_{it})'$. Of course, with some redundancy, we then have $\mathbf{Y}_i \equiv \mathbf{Y}_{iJ}$ and $\mathbf{Y}_{i,\text{obs}} \equiv \mathbf{Y}_{iT_i}$.

We express (1) in terms of the odds that $Y_{it} = 1$. As such, let $\theta_{it} = \theta_i(X_{it}; \beta) = \Pr(Y_{it} = 1|\mathbf{X}_i)/\Pr(Y_{it} = 0|\mathbf{X}_i)$. The goal is to make inferences about the $p \times 1$ parameter β in the logistic model given by $\log(\theta_{it}) = q_i + X_{it}'\beta$, where q_i is a subject-specific intercept which allows $\Pr(Y_{it} = 1)$ to vary across

subjects according to unobserved subject-level variables. Note that covariates Z_i do not figure into this model, as the effects of these covariates are absorbed into the intercept q_i.

2.2. Drop-out model

To account for the drop-out process, define

$$\begin{aligned}\lambda_i(t, \mathbf{y}_{t-1}; \gamma) &= \Pr(R_{it} = 1 | R_{i1} = \cdots = R_{i,t-1} = 1, \mathbf{Y}_i = \mathbf{y}, \mathbf{X}_i, Z_i) \\ &= \Pr(R_{it} = 1 | R_{i,t-1} = 1, \mathbf{Y}_i = \mathbf{y}, \mathbf{X}_i, Z_i),\end{aligned}$$

where γ is a finite-dimensional nuisance parameter which does not depend on response model parameters (q_i, β), $\mathbf{y} = (y_1, \ldots, y_J)'$, and the role of subscript $t-1$ on the bold $\mathbf{y}$ is as defined above. Note that $1 - \lambda_i(t, \mathbf{y}_{t-1})$ is the hazard of drop-out at t, and that (2) ensures that $\lambda_i(t, \mathbf{Y}_{i,t-1})$ does not depend on $Y_{it}, \ldots, Y_{iJ}$. For example, $\lambda_i(t, \mathbf{Y}_{i,t-1})$ might depend on past values $Y_{it'}, t' < t$, via a logistic regression model that only depends on the most recent value $Y_{i,t-1}$ independently of t. Alternatively, $\lambda_i(t, \mathbf{Y}_{i,t-1})$ might vary across t and/or depend on more than just the most recent value of $Y_{it'}, t < t'$. Also, $\lambda_i(t, \mathbf{Y}_{i,t-1})$ is implicitly allowed to depend on subject-level covariates Z_i, such as sex or treatment assignment, as well as the matrix $\mathbf{X}_i$ of time-varying covariates, which might include time or treatment-by-time interactions. Throughout, we assume that dependence of $\lambda_i(t, \mathbf{y}_{t-1})$ on $(\mathbf{X}_i, Z_i)$ is indicated by subscript i, but we make the dependence on $\mathbf{y}_{t-1}$ explicit for reasons that will become clear. We also assume that $\lambda_i(1, \mathbf{y}_0) \equiv 1$, i.e., the baseline assessment is always observed, and, for ease of exposition, that $\lambda_i(J+1, \mathbf{y}_J) \equiv 0$.

Given a model for $\lambda_i(t, \mathbf{y}_{t-1})$, the drop-out probability is immediately computed as

$$\pi_i(t, \mathbf{y}_t; \gamma) = \Pr(T_i = t | \mathbf{Y}_i = \mathbf{y}, \mathbf{X}_i, Z_i) = \prod_{s=1}^{t} \lambda_i(s, \mathbf{y}_{s-1}; \gamma)\{1 - \lambda_i(t+1, \mathbf{y}_t; \gamma)\}.$$

By (2), $\pi_i(t, \mathbf{Y}_{it})$ depends only on data observed at or before t.

2.3. Complete-record conditional likelihood analysis

With the models defined in the previous section, we are now in a position to define the likelihood arising from the complete-record data $Y_{i1}, \ldots, Y_{iT_i}, \mathbf{X}_i, Z_i$. This likelihood, L_i, is conditional on the drop-out process T_i, that is, $L_i(\beta, \gamma, q_i) = \Pr(\mathbf{Y}_{i,\text{obs}} | \mathbf{X}_i, Z_i, T_i)$. Expressing this likelihood via odds θ_{it} and drop-out probability $\pi_i(t, \mathbf{y}_t)$, yields

$$L_i(\beta, \gamma, q_i) = \frac{\{\prod_{t=1}^{T_i} \theta_{it}^{Y_{it}}\} \pi_i(T_i, \mathbf{Y}_{i,\text{obs}})}{\sum_{\mathbf{y}_{\text{obs}} \in \mathcal{Y}^*_{i,\text{obs}}} \{\prod_{t=1}^{T_i} \theta_{it}^{y_t}\} \pi_i(T_i, \mathbf{y}_{\text{obs}})}, \tag{3}$$

where $\mathcal{Y}^*_{i,\text{obs}}$ is the set of all 2^{T_i} possible vectors $\mathbf{y}_{\text{obs}} = (y_1, \ldots, y_{T_i})'$.

The difficulty with using L_i for inferences about β is the presence of nuisance parameters q_i and γ. First, consider the subject intercept q_i. Via standard theory of exponential family models, it is easily seen that, for fixed β and γ, $\sum_{t=1}^{T_i} Y_{it}$ is a complete sufficient statistic for q_i in (3). Therefore, conditioning on this statistic will yield a likelihood that is free of q_i. Let

$$L_i^c(\beta, \gamma) = \Pr(\mathbf{Y}_{i,\text{obs}} | \textstyle\sum_{t=1}^{T_i} Y_{it}, \mathbf{X}_i, Z_i, T_i). \tag{4}$$

Then

$$L_i^c(\beta, \gamma) = \frac{\{\prod_{t=1}^{T_i} (e^{X_{it}'\beta})^{Y_{it}}\} \pi_i(T_i, \mathbf{Y}_{i,\text{obs}})}{\sum_{\mathbf{y}_{\text{obs}} \in \mathcal{Y}_{i,\text{obs}}} \{\prod_{t=1}^{T_i} (e^{X_{it}'\beta})^{y_t}\} \pi_i(T_i, \mathbf{y}_{\text{obs}})}\ , \tag{5}$$

where $\mathcal{Y}_{i,\text{obs}} = \mathcal{Y}_{i,\text{obs}}(\mathbf{Y}_{i,\text{obs}})$ is the set of all $\binom{T_i}{\sum_{i=1}^{T_i} Y_{it}}$ vectors $\mathbf{y}_{\text{obs}} = (y_1, \ldots, y_{T_i})'$ such that $\sum_{t=1}^{T_i} y_t = \sum_{t=1}^{T_i} Y_{it}$.

Remark. Likelihood (5) can be contrasted with the standard conditional likelihood which ignores the drop-out process. This standard likelihood deletes the $\pi_i(\cdot)$ terms from (5) and is equivalent to assuming that $\pi_i(T_i, \mathbf{y}_{\text{obs}})$ is constant across all $\mathbf{y}_{\text{obs}} \in \mathcal{Y}_{i,\text{obs}}$. The implication is that the standard conditional likelihood method is biased under MAR. The fact that one must account for the drop-out process is surprising because in likelihood-based approaches, MAR dropout processes are generally assumed to be ignorable. Further elaboration on this point is given in the discussion.

For estimation of γ, we model R_{it} among those subjects for whom $R_{i,t-1} = 1$. As such, define L_i^γ to be the ith subject's contribution to the γ-likelihood, i.e.,

$$L_i^\gamma(\gamma) = \Big\{ \prod_{t=1}^{T_i - 1} \lambda_i(t, \mathbf{Y}_{i,t-1}; \gamma) \Big\} \{1 - \lambda_i(T_i, \mathbf{Y}_{i,T_i-1}; \gamma)\}.$$

Then, accumulating information over subjects $i = 1, \ldots, K$, let $\widehat{\gamma}$ be the estimator of γ obtained by maximizing the likelihood $\prod_i L_i^\gamma(\gamma)$.

We propose maximum conditional likelihood estimation of β using L_i^c, with $\widehat{\gamma}$ replacing γ. Combining information across subjects, let $\widehat{\beta}$ be the maximizer of $\prod_i L_i^c(\beta, \widehat{\gamma})$. Equivalently, $\widehat{\beta}$ is the solution to $\sum_i U_i^c(\beta, \widehat{\gamma}) = 0$, where $U_i^c(\beta, \gamma)$ is the ith subject's β-score contribution $U_i^c = (\partial \log L_i^c / \partial \beta)$, $\widehat{\gamma}$ is the solution to $\sum_i S_i^\gamma(\gamma) = 0$, and $S_i^\gamma(\gamma)$ is the ith subject's γ-score contribution $S_i^\gamma = (\partial \log L_i^\gamma / \partial \gamma)$. An estimator for the asymptotic variance of $\widehat{\beta}$ with estimated γ is given as a special case of Theorem 1 by replacing U_i there with U_i^c.

Remark. In the special case where $\lambda_i(t, \mathbf{Y}_{i,t-1})$ depends only on the most recent response value, that is, $\lambda_i(t, \mathbf{y}_{t-1}) = \lambda_i(t, y_{t-1})$, $\widehat{\beta}$ can be computed using standard software, as follows. First define

$$B_{it} = \begin{cases} \log\{\lambda_i(t+1,1)/\lambda_i(t+1,0)\}, & t = 1, \ldots, T_i - 1 \\ \log[\{1-\lambda_i(t+1,1)\}/\{1-\lambda_i(t+1,0)\}], & t = T_i. \end{cases}$$

It is easily shown that

$$L_i^c(\beta, \gamma) = \frac{\prod_{t=1}^{T_i} (e^{X_{it}'\beta + B_{it}})^{Y_{it}}}{\sum_{\mathbf{y}_{\text{obs}} \in \mathcal{Y}_{i,\text{obs}}} \prod_{t=1}^{T_i} (e^{X_{it}'\beta + B_{it}}) y_t} \ , \tag{6}$$

so that the model accounting for drop-out can be fitted using a standard conditional logistic regression software package, including the offset B_{it} in the linear predictor. The resulting estimator for β will be consistent, although the standard errors produced by the package will be conservative.

3. Efficiency Improvements

3.1. Introduction

While the conditional likelihood L_i^c will yield valid inferences about β, it is potentially inefficient because it contains information on the missingness process $\mathbf{R}_i$, which is ancillary for β. To see this, note that likelihood L_i^γ depends in no way on β, and yet L_i^c depends on random variables $\mathbf{R}_i$. This suggests that more efficient estimation of β can be achieved by using an estimating function where the β-ancillary information in $\mathbf{R}_i$ has been removed. Heuristically, the idea is to identify an estimating function U_i^a that (i) contains no β-information, i.e., is ancillary for β, (ii) is unbiased without any further modeling assumptions, and (iii) is positively correlated with U_i^c. Here, we take U_i^a being 'ancillary for β' to mean that $\mathrm{E}\,(-\partial U_i^a/\partial\beta) = 0$. Then, $U_i^c - U_i^a$ will yield a potential increase in efficiency relative to U_i^c.

At (7) and (8), we introduce a class $\mathcal{U}_i$ of estimating functions for β that includes the complete-record estimating function U_i^c and that satisfies criteria (i) and (ii) above. Motivation for the general form of this class, and its component functions $V_i^{(t,s)}$, is elaborated in Section 3.3, where we use a projection argument and semiparametric efficiency theory to generate the optimal member U_i^{proj} of $\mathcal{U}_i$. We show that U_i^{proj} satisfies (iii) and yields more efficient β inferences than U_i^c. Finally, in Section 3.4, we propose an approximation U_i^{appr} to U_i^{proj}, also in $\mathcal{U}_i$, that is more practical to compute than U_i^{proj} and that is also expected to satisfy (iii). Simulations in Section 4 illustrate the potential efficiency improvements in U_i^{appr} over U_i^c.

Throughout Section 3, we treat q_i as an unobserved random variable with a nonparametric distribution depending arbitrarily on $(\mathbf{X}_i, Z_i)$. We note that conditional likelihood L_i^c and scores U_i^c remain valid under this model. References to the joint distribution of $\mathbf{Y}_i$ or of its sub-vectors $\mathbf{Y}_{it}$ are conditional on $(\mathbf{X}_i, Z_i)$, but marginal over q_i. That is, $(\mathbf{Y}_i|\mathbf{X}, Z_i)$ has a nonparametric mixture distribution. Throughout Section 3, detailed proofs and technical material are omitted; further information is available in a technical report from the author. Finally, as most development in this section is at the subject level, the subscript i is omitted except where needed for clarity.

3.2. A class of estimating functions

First, for a given subject i and for every $(t, s), 1 \leq s \leq t+1, 1 \leq t \leq J$, define arbitrary functions $V^{(t,s)}(R_s, \mathbf{Y}_{s-1}, \mathbf{X}, Z; \beta, \gamma, \alpha)$. Generally, $V^{(t,s)}$ will depend on β, γ and possibly a finite dimensional nuisance parameter α. Set $V^{(J,J+1)} \equiv 0$.

Now, consider estimating functions of the form

$$U^a = \sum_{t=1}^{J} \sum_{s=1}^{t+1} R_{s-1}(V^{(t,s)} - \epsilon^{(t,s)}), \tag{7}$$

where we define $R_0 \equiv 1$ and, taking expectation over R_s, $\epsilon^{(t,s)} = \mathrm{E}\,(V^{(t,s)}|R_{s-1} = 1, \mathbf{Y}_{s-1}, \mathbf{X}, Z)$. Note that, owing to the factor R_{s-1} in (7), U^a can always be computed with observed data. It is straightforward to show that, regardless of choice of $V^{(t,s)}$, $\mathrm{E}\,(U^a) = \mathrm{E}\,(-\partial U^a/\partial\beta) = 0$ as long as the missingness model $\lambda_i(t, \mathbf{y}_{t-1})$ is correctly-specified, so that (7) defines a class of unbiased β-ancillary estimating functions U^a satisfying criteria (i) and (ii) in Section 3.1, and indexed by the choice of functions $V^{(t,s)}$. The elements of (7) yield a class $\mathcal{U}$ of estimating functions for β of the form

$$U(\beta, \gamma, \alpha) = U^c(\beta, \gamma) - U^a(\beta, \gamma, \alpha). \tag{8}$$

Specific members of $\mathcal{U}$ are the subjects of Sections 3.3−3.4.

For a given choice of $V^{(t,s)}$, to use the resulting estimating function $U \in \mathcal{U}$ for β-inferences, assume that there exists an α-estimating function $S^\alpha = S^\alpha(\mathbf{Y}_{\mathrm{obs}}, \mathbf{X}_{\mathrm{obs}}, Z, T; \alpha)$, and let $\widehat{\alpha}$ be the solution to $\sum_i S_i^\alpha(\alpha) = 0$. Suppose that we estimate γ as in Section 2.3, and β by solving $\sum_i U_i(\beta, \widehat{\gamma}, \widehat{\alpha}) = 0$. Theorem 1 characterizes the asymptotic distribution of $\widehat{\beta}$. A sketch of a proof is in Appendix A.

Theorem 1. *Suppose that $\widehat{\gamma}$ solving $\sum_i S_i^\gamma(\gamma) = 0$ is $\sqrt{K}$-consistent, that $\widehat{\alpha}$ is $\sqrt{K}$-consistent for some α^*, and that $\widehat{\beta}$ solves $\sum_i U_i(\beta, \widehat{\gamma}, \widehat{\alpha}) = 0$. Then, under*

mild regularity conditions as $K \to \infty$, $\widehat{\beta} \to \beta$ *in probability, and* $\sqrt{K}(\widehat{\beta} - \beta) \to N(0, \mathcal{V})$ *in distribution,* $\mathcal{V} = \lim_{K\to\infty} \mathcal{V}_K$, *where*

$$\mathcal{V}_K = K(\sum_i \mathcal{I}^c_{\beta i})^{-1}(\sum_i \widetilde{U}_i \widetilde{U}_i^{\mathrm{T}})(\sum_i \mathcal{I}^c_{\beta i})^{-1}, \tag{9}$$

$\mathcal{I}^c_{\beta i} = \mathrm{E}\,(U^c_i U^{c\,\mathrm{T}}_i)$, *and* $\widetilde{U}_i = U_i - C S^\gamma_i$, *where* $C = (\sum_i U_i S^{\gamma\,\mathrm{T}}_i)\{\sum_i S^\gamma_i S^{\gamma\,\mathrm{T}}_i\}^{-1}$.

Remark. From (9), we see that the asymptotic efficiency of $\widehat{\beta}$ for U^c or any $U \in \mathcal{U}$ is improved by estimation of γ, even if γ is already known. Similarly, using a model for the drop-out process $\lambda(t, \mathbf{y}_{t-1})$ that is richer than required, for example by including unnecessary interaction terms, will not harm efficiency and may yield further gains. These phenomena have been previously noted in missing data problems (Robins, Rotnitzky and Zhao (1995)). We explore their implications for a simple model in the simulations presented in Section 4.

3.3. Improved efficiency via projection

We now exploit ideas in semiparametric efficiency theory to obtain a member U^{proj} that is optimal in the class $\mathcal{U}$. The key idea is to remove from U^c its projection onto the tangent space $\mathcal{W}$ for the nuisance parameter γ (i.e., the closed linear span of $\mathcal{L}^2$ scores for γ; Newey (1990, Section 3) and Robins, Rotnitzky and van der Laan (2000, Section 3)). To do so, we first establish a representation of $\mathcal{W}$. Then, we rewrite U^c as a sum over all possible drop-out times, which facilitates computing the projection.

We show in a technical report that $\mathcal{W}$ is the $\mathcal{L}^2$ subspace spanned by the union of subspaces $\mathcal{W}_s$, indexed by $s = 1, \ldots, J$, where $\mathcal{W}_s$ is the $\mathcal{L}^2$ subspace of functions of $(R_s, R_{s-1}, \mathbf{Y}_{s-1}, \mathbf{X}, Z)$ which are unbiased conditional on $(R_{s-1}, \mathbf{Y}_{s-1}, \mathbf{X}, Z)$. Let $\mathcal{P}$ ($\mathcal{P}_s$) be the $\mathcal{L}^2$ projection operator into $\mathcal{W}$ ($\mathcal{W}_s$). By the definition of $\mathcal{W}_s$, for any $\mathcal{L}^2$ regular estimating function g,

$$\mathcal{P}_s g = \mathrm{E}\,(g|R_s, R_{s-1}, \mathbf{Y}_{s-1}, \mathbf{X}, Z) - \mathrm{E}\,(g|R_{s-1}, \mathbf{Y}_{s-1}, \mathbf{X}, Z).$$

Also, because the $\mathcal{W}_s$'s are orthogonal to one another, the projection of g onto $\mathcal{W}$ is just the sum of the projections onto the $\mathcal{W}_s$'s, that is, $\mathcal{P}g = \sum_{s=1}^J \mathcal{P}_s g$.

Further development requires that the dependence of U^c on T (or $\mathbf{R}$) be explicit. As such, we write a version of conditional likelihood L^c corresponding to each of the $t = 1, \ldots, J$ possible drop-out times. Let $L^{(t)}$ be the value that L^c takes when $T = t$, that is, following (4), $L^{(t)}(\beta, \gamma) = \Pr(\mathbf{Y}_t | \sum_{s=1}^t Y_s, \mathbf{X}, Z, T = t)$. Then, $L^c = \prod_{t=t}^J L^{(t)\,I(T=t)}$. Similarly, writing $U^{(t)} = (\partial \log L^{(t)}/\partial\beta)$, we can rewrite U^c as

$$U^c = \sum_{t=1}^J I(T = t) U^{(t)}. \tag{10}$$

To compute projections $\mathcal{P}_s U^c$, we exploit (10), operating one term at a time. It can be shown that, for $1 \le s \le t+1$, $1 \le t \le J$,

$$\mathcal{P}_s I(T=t)U^{(t)} = R_{s-1}\left(V_{\text{proj}}^{(t,s)} - \epsilon_{\text{proj}}^{(t,s)}\right), \tag{11}$$

where $V_{\text{proj}}^{(t,s)}(\beta,\gamma,\alpha) = \mathrm{E}\{(1-R_{t+1})R_t U^{(t)} | R_s, R_{s-1}=1, \mathbf{Y}_{s-1}, \mathbf{X}, Z\}$, expectation being taken over $(\mathbf{Y},\mathbf{R})$, and $\epsilon_{\text{proj}}^{(t,s)}(\beta,\gamma,\alpha) = \mathrm{E}(V_{\text{proj}}^{(t,s)} | R_{s-1}=1, \mathbf{Y}_{s-1}, \mathbf{X}, Z)$, expectation being over R_s. Similarly, for $s > t+1$, $\mathcal{P}_s I(T=t)U^{(t)} = 0$ and, for ease of exposition, we define $\mathcal{P}_{J+1} g \equiv 0$.

Summing over (t,s), the projection of U^c onto γ-tangent space $\mathcal{W}$ is therefore

$$\mathcal{P}U^c = \sum_{t=1}^{J}\sum_{s=1}^{t+1} R_{s-1}\left(V_{\text{proj}}^{(t,s)} - \epsilon_{\text{proj}}^{(t,s)}\right) \tag{12}$$

and, subtracting $\mathcal{P}U^c$ from U^c, we thereby define a new estimating function

$$U^{\text{proj}} = U^c - \mathcal{P}U^c = U^c - \sum_{t=1}^{J}\sum_{s=1}^{t+1} R_{s-1}\left(V_{\text{proj}}^{(t,s)} - \epsilon_{\text{proj}}^{(t,s)}\right)$$

for inferences about β. Since $V_{\text{proj}}^{(t,s)}$ is a function of $(R_s, \mathbf{Y}_{s-1}, \mathbf{X}, Z)$, $\mathcal{P}U^c$ satisfies (7), and U^{proj} is therefore in the class $\mathcal{U}$ defined by (8). Theorem 2 elucidates the efficiency benefit in using U^{proj} over U^c, or any other $U \in \mathcal{U}$, for inferences about β, and implies that projection (12) satisfies criterion (iii) in Section 3.1.

Theorem 2. *U^{proj} is optimally efficient in $\mathcal{U}$ in the sense that, for any $U \in \mathcal{U}$, $\mathcal{I}_\beta^{\text{proj}} - \mathcal{I}_\beta$ is positive semidefinite, where the U^{proj} information matrix $\mathcal{I}_\beta^{\text{proj}}$ is $\mathcal{I}_\beta^{\text{proj}} = \mathcal{I}_\beta^c \mathrm{E}(U^{\text{proj}} U^{\text{proj}\,\mathrm{T}})^{-1} \mathcal{I}_\beta^c$ and $\mathcal{I}_\beta$ is the corresponding information matrix for U.*

Sketch of a proof. The proof of Theorem 2 involves three main points, which hold for any $U \in \mathcal{U}$. First, $\mathrm{E}(-\partial U/\partial\beta) = \mathcal{I}_\beta^c$. Second, $U - \mathcal{P}U = U^c - \mathcal{P}U^c = U^{\text{proj}}$. Third, $\mathcal{P}U$ is positively-correlated with U, so that

$$\mathrm{E}(UU^{\mathrm{T}}) - \mathrm{E}(U^{\text{proj}} U^{\text{proj}\,\mathrm{T}}) \ge 0 \tag{13}$$

in the positive semidefinite sense.

Remarks.

1. Theorem 1 demonstrated, for a given choice $U \in \mathcal{U}$, that efficiency is improved by estimation of γ and by using overly-rich models for the drop-out process. However, it said nothing about the choice of U^a (i.e., of $V^{(t,s)}$) in $U = U^c - U^a$. Theorem 2 provides guidance in choosing U^a to improve efficiency.

2. Note that $\mathcal{P}U^c$, and hence U^{proj}, depend on the joint distribution of $(\mathbf{Y}|\mathbf{X}, Z)$ to compute $V^{(t,s)}_{\text{proj}}$, and the distribution of $(\mathbf{Y}|\mathbf{X}, Z)$ in turn depends on a model for the mixture distribution of $(q|\mathbf{X}, Z)$. This is the first place in our development where $(q|\mathbf{X}, Z)$ is needed. However, note that if the model for $(\mathbf{Y}|\mathbf{X}, Z)$ is misspecified, the resulting $V^{(t,s)}_{\text{proj}}$'s will not be the optimal functions, but they will still be valid functions $V^{(t,s)}$ as defined in Section 3.2. Also, given any $V^{(t,s)}$ ($V^{(t,s)}_{\text{proj}}$), correct computation of $\epsilon^{(t,s)}$ ($\epsilon^{(t,s)}_{\text{proj}}$) only depends on the drop-out model, $\lambda(t; \mathbf{y}_{t-1})$. The implication is that even if the joint distribution of $(\mathbf{Y}|\mathbf{X}, Z)$ is misspecified, the resulting $\mathcal{P}U^c$, while not optimal, will still be a valid member of the class (7), and the resulting $U^{\text{proj}} = U^c - \mathcal{P}U^c$ will still be a valid unbiased estimating function in the class $\mathcal{U}$. The main assumption required throughout our development of U^c, U^{proj} and the class $\mathcal{U}$ is that the drop-out model (3) be correctly specified.

3. Even if the joint distribution $(\mathbf{Y}|\mathbf{X}, Z)$ and resulting $\mathcal{P}U^c$ are only approximately correct, we still might expect an increase in efficiency in U^{proj} over U^c. The reason is that, even an approximate $\mathcal{P}U^c$ will be positively correlated with U^c (criteria (iii)), so that (13) in the sketch of proof will still hold.

3.4. A practical estimator

The results of the foregoing section show that, to compute the $V^{(t,s)}_{\text{proj}}$'s and the projection $\mathcal{P}U^c$ needed for U^{proj}, we require the the full joint distribution $(\mathbf{Y}|\mathbf{X}, Z)$, marginally over q. In computing U^{proj}, distribution $(\mathbf{Y}|\mathbf{X}, Z)$ plays a role in increasing efficiency of $\widehat{\beta}$, but correct specification of this distribution is not critical for consistency. Also, it may be difficult to compute the $V^{(t,s)}_{\text{proj}}$'s, as they require specification and estimation of the unknown mixture distribution $(q|\mathbf{X}, Z)$, and then complicated numerical integration over this distribution. Indeed, avoiding specification of $(q|\mathbf{X}, Z)$ was one motivation for using a fixed effects model in the first place. Therefore, in real data-analytic settings, it may be practical to employ a working model for $(\mathbf{Y}|\mathbf{X}, Z)$ for purposes of computing $\mathcal{P}U^c$. In this section, we accomplish this by using a parametric working model for $(\mathbf{Y}|\mathbf{X}, Z)$ based on the transition distribution of $(Y_t|\mathbf{Y}_{t-1}, \mathbf{X}, Z)$. We emphasize that (1) is still assumed to be the true model, and that the working transition model is only to be used for approximating the projection operator $\mathcal{P}$ that will be applied to U^c. Specifically, the working transition model is used to obtain approximations $V^{(t,s)}_{\text{appr}}$ to $V^{(t,s)}_{\text{proj}}$; $\epsilon^{(t,s)}_{\text{appr}}$ is then obtained from $V^{(t,s)}_{\text{appr}}$ as in (7).

Define the transition probabilities

$$\eta(t, \mathbf{y}_{t-1}; \alpha) = \Pr(Y_t = 1|\mathbf{Y}_{t-1} = \mathbf{y}_{t-1}, \mathbf{X}, Z), \tag{14}$$

for $t = 2, \ldots, J$, where α is a finite-dimensional nuisance parameter. In principle, model $\eta(\cdot; \alpha)$ and parameter α depend on the interest parameter β. However, exploiting this dependency requires specification of the mixture distribution

$(q|\mathbf{X}, Z)$ which we would like to avoid. Even though it ignores these dependencies (14), meets our needs for a practical approximation to distribution $(\mathbf{Y}|\mathbf{X}, Z)$.

Model (14) for $(\mathbf{Y}|\mathbf{X}, Z)$, together with the model $\lambda(t, \mathbf{y}_{t-1}; \gamma)$ for the dropout process $(\mathbf{R}|\mathbf{Y}, \mathbf{X}, Z)$, yields a working model for the joint distribution of $(\mathbf{R}, \mathbf{Y}|\mathbf{X}, Z)$. This model is used in place of the true but unknown distribution to approximate the projection $\mathcal{P}U^c$. As such, define functions $V_{\text{appr}}^{(t,s)}$ as

$$V_{\text{appr}}^{(t,s)}(\gamma, \alpha) = \widetilde{\mathrm{E}}\left\{(1 - R_{t+1})R_t U^{(t)} | R_s, R_{s-1} = 1, \mathbf{Y}_{s-1}, \mathbf{X}, Z; \alpha, \gamma\right\},$$

where $\widetilde{\mathrm{E}}(\cdot; \alpha, \gamma)$ denotes expectation taken with respect to the working distribution of $(\mathbf{R}, \mathbf{Y}|\mathbf{X}, Z)$. Note importantly that functions $V_{\text{appr}}^{(t,s)}$ are of the form $V^{(t,s)}$ given in Section 3.2, and hence they define an element U^{appr} in $\mathcal{U}$. Specifically,

$$U^{\text{appr}}(\beta, \gamma, \alpha) = U^c - \sum_{t=1}^{J} \sum_{s=1}^{t+1} R_{s-1}(V_{\text{appr}}^{(t,s)} - \epsilon_{\text{appr}}^{(t,s)}).$$

We propose U^{appr} as an improvement over U^c for inferences on β.

In order to use U^{appr}, we require an estimator of the transition model parameter α. For this, note that, under (2), the ith subject's contribution to the likelihood function L_i^α for α is given by

$$L_i^\alpha(\alpha) = \prod_{t=2}^{J} [\eta_i(t, \mathbf{Y}_{i,t-1}; \alpha)^{Y_{it}} \{1 - \eta_i(t, \mathbf{Y}_{i,t-1}; \alpha)^{(1-Y_{it})}\}]^{R_{it}}.$$

Accumulating information over subjects $i = 1, \ldots, K$, let $\widehat{\alpha}$ be the estimator of α obtained by maximizing the likelihood $\prod_i L_i^\alpha(\alpha)$. We assume that $\widehat{\alpha}$ is consistent for some value α^* even if (14) is not valid or is misspecified. Note that, similarly to $\widehat{\gamma}$, $\widehat{\alpha}$ can be computed via L_i^α before estimation of β. With estimators $\widehat{\gamma}$ and $\widehat{\alpha}$, $U^{\text{appr}}(\beta, \widehat{\gamma}, \widehat{\alpha})$ is an approximate projected estimating function, and the resulting estimator $\widehat{\beta}$ solving $\sum_i U_i^{\text{appr}}(\beta, \widehat{\gamma}, \widehat{\alpha}) = 0$ has the asymptotic distribution given in Theorem 1. We investigate the performance of U^{appr} compared to U^c via simulations, in the next section.

4. Simulation Study

To investigate the performance of the estimators based on U^c and U^{appr}, as compared with those arising from the standard conditional likelihood based on the first T_i records, we performed a small simulation study. Each replicate sample consisted of $i = 1, \ldots, K = 200$ subjects potentially measured at $t = 1, \ldots, J = 5$ equally-spaced times. Each subject was randomly assigned treatment $Z_i = 1$ or control $Z_i = 0$, with about 31% of subjects receiving $Z_i = 1$. The model that was used both to generate and to analyze the data was $\text{logit}\{\Pr(Y_{it} = 1|\mathbf{X}_i, q_i)\} =$

$q_i + \beta_t(t-1)/4 + \beta_x X_{it}$, where $X_{it} = Z_i \times (t-1)/4$. Note that $(t-1)/4$ ranges from zero to one. We set $\beta_t = \beta_x = \log(1.5)$ and $q_i = \{(i-1)/199\}^2 - 1.5$, yielding a marginal $\Pr(Y_{it} = 1) = 0.30$. Drop-out was generated using the model

$$\text{logit}\{\Pr(R_{it}=1|R_{i,t-1}=1, \mathbf{Y}_{i,t-1}, \mathbf{X}_i, Z_i)\} = \gamma_0 + \gamma_t(t-1)/4 + \gamma_y Y_{i,t-1} + \gamma_z Z_i \quad (15)$$

for $t = 2, \ldots, 5$, with this probability being one for $t = 1$ and zero for $t = 6$. We set $\gamma = (\gamma_0, \gamma_t, \gamma_y, \gamma_z) = (1.6, 0.1, 0.4, 0.4)^T$, yielding a drop-out hazard of 16−21% across $t = 1, \ldots, 4$, and 44% of subjects with complete data through $t = 5$.

For each replicate, seven estimators were computed. The first was the estimator computed using standard CLR applied to the first T_i observed records for each subject i. The next three estimators were based on the bias-corrected likelihood L_i^c and score U_i^c proposed in Section 2.3. For the second estimator, we assumed that $\lambda_i(t; \mathbf{y}_{t-1})$, and hence $\pi_i(t, \mathbf{y}_{t-1})$, were known. In the third, we estimated $\lambda_i(t; \mathbf{y}_{t-1})$ using (15), performing maximum likelihood estimation with L_i^γ to obtain $\widehat{\gamma}$, as described in Section 2.3. We refer to (15) as the 'minimal' drop-out model. The fourth estimator used a richer drop-out model than required, adding all two- and three-way interactions between t, $Y_{i,t-1}$ and Z_i to (15). We call this the 'rich' drop-out model. We include it to evaluate potential increases in efficiency due to over-specification of the missingness model. Finally, three estimators based on U^{appr} were computed. The first used a 'minimal' transition model, modeling Y_{it} as a function of $Y_{i,t-1}$ only. Here, (15) was used. In the second and third, a 'full' transition model was used, modeling $\text{logit}\{\Pr(Y_{it} = 1|\mathbf{Y}_{i,t-1}, \mathbf{X}_i, Z_i, q_i)\} = \gamma_0 + \alpha_t(t-1)/4 + \alpha_y Y_{i,t-1} + \alpha_z Z_i$ for $t = 2, \ldots, 5$. In the third estimator, the 'rich' drop-out model was used. For all estimators, we computed 95% Wald-type confidence intervals based on the variance estimator (9).

Results based on 1,000 replicates are reported in Table 1. These include percent bias, mean square error efficiency relative to the bias-corrected estimator with estimated γ under the 'minimal' drop-out model, and confidence interval coverage probabilities. As can be seen, the standard CLR estimator in this setting is strongly biased. All of the new estimators correct this bias. When using the bias-corrected U^c, estimation of γ improves efficiency in $\widehat{\beta}$ relative to the case where γ is assumed known, and using a richer drop-out model than required yields an additional 5−10% efficiency improvement. In contrast, the approximate projection method using U^{appr} yields a 15−20% efficiency improvement. This efficiency gain is robust to the transition model chosen for $(Y_{it}|Y_{i,t-1}, X_{it}, Z_i)$, as the results are the same for the 'minimal' and the 'full' transition models. Finally, when using U^{appr}, richer drop-out models no longer provide efficiency

improvements, as the projection $\mathcal{P}U^c$ has effectively already accounted for all such improvements.

Table 1. Simulation results based on 1,000 replicates. Upper entries are for β_t and lower entries are for β_x. True values are $\beta_t = \beta_x = 0.405$.

Method	Drop-out Model	Mean($\hat{\beta}$)	% Bias	SE($\hat{\beta}$)	Rel. Eff.	Cov. %
Standard complete case	–	0.230 0.453	-43.2 11.6	0.350 0.570	74 100	91.3 94.7
U^c, known λ	'min'	0.401 0.405	-1.0 -0.1	0.354 0.574	90 99	95.4 94.9
U^c, est. λ	'min'	0.401 0.405	-1.2 -0.1	0.336 0.570	– –	95.1 95.1
U^c, est. λ	'rich'	0.401 0.407	-1.1 0.5	0.328 0.542	105 111	95.3 95.0
U^{appr}, 'min' Y_t model	'min'	0.400 0.405	-1.3 0.0	0.313 0.522	115 119	95.3 94.6
U^{appr}, 'full' Y_t model	'min'	0.400 0.405	-1.3 -0.1	0.313 0.521	115 120	95.6 94.5
U^{appr}, 'full' Y_t model	'rich'	0.401 0.405	-1.1 -0.1	0.313 0.522	115 120	95.2 94.5

Rel. Eff., mean square error efficiency relative to U^c with estimated 'minimal' drop-out model. Cov. %, coverage percent for 95% Wald-type confidence intervals.

5. Disability–Memory Example

We now revisit the aging research example introduced in Section 1. In that study, baseline (year 0) data were collected in 1993, and follow up data were collected two, five and seven years later. Disability here is a assessed via a binary variable indicating whether the subject reports difficulty with at least one of the following activities: preparing hot meals, shopping for groceries, making telephone calls, taking medications and managing money. Memory was assessed at baseline by the sum of immediate and delayed word recall. A list of ten words was read aloud and the respondent was asked to repeat as many as possible; the resulting count of correct words is the immediate word recall. About five minutes later, after some other questions, the respondent was again asked to name as many of the words as he or she remembered, providing a measure of delayed word recall. The average of the two, scaled to have standard deviation one, is used in this analysis.

Interest is on the change in disability over time, and how that change is affected by level of memory at baseline. We adjust our results for age, sex and education, all measured at baseline. Because we are interested in change, the key parameters of interest are the interactions between year and each of age, sex, education and memory. We model the main effect of year non-parametrically, with a dummy variable for each year of follow up (Table 2). This allows for non-linear effects of time on study which, in fact, we observe in the data. Deviations from this trend are modelled smoothly, however, through interactions of covariates with linear year. Our first analysis uses standard conditional logistic regression on the available data, yielding the estimates of subject-specific log odds ratios in the first column of Table 2.

Table 2. Fixed effects models for changes in disability as a function of baseline factors.

	Standard	Bias-corrected		Efficiency-improved	
	Est.	Est.	SE	Est.	SE
$I(\text{Year} = 2)$	-0.030	-0.20	(0.07)	-0.16	(0.07)
$I(\text{Year} = 5)$	0.92	0.66	(0.11)	0.63	(0.10)
$I(\text{Year} = 7)$	1.44	1.04	(0.14)	0.95	(0.14)
Age × year	0.11	0.095	(0.021)	0.096	(0.020)
Sex × year	0.047	0.058	(0.022)	0.064	(0.022)
Education × year	0.059	0.045	(0.012)	0.045	(0.012)
Memory × year	-0.033	-0.028	(0.012)	-0.024	(0.012)

Est., $\widehat{\beta}$ parameter estimates of subject-specific log odds ratios.
SE, robust standard errors correcting for estimation of drop-out model.
Covariates: Baseline year is 0. Age is in 10-year units, centered at 80 years. Education is in four year units, centered at 12 years. Memory has mean zero, standard deviation one.

As mentioned earlier, attrition is a major concern in this analysis. We fitted a logistic drop-out model $\lambda_i(t;\cdot)$ at waves $t = 2, 3, 4$ that contained main effects of year, age, sex, education, memory and disability at time $t-1$, as well as interactions between year and age, sex, education and disability, and between education and disability. Other two-way interaction terms contributed very little to the fit of the dropout model. We assume here that the drop-out yields data that are missing at random. We note that it may very well be true that subject drop-out is associated with disability response Y_{it} for t after drop-out time T_i. However, the MAR assumption does not require that these processes be marginally independent; rather, MAR only requires the milder assumption that they be independent given covariates in the model and all observed disability outcomes up to drop-out. As drop-out due to death, serious decline in health

and/or a change in residence is likely to be preceded by an increase in disability as measured by the ability to due complex tasks such as grocery shopping or managing money, this assumption does not appear unreasonable.

Using this drop-out model with U^c and the method in Section 2.3, bias-corrected estimates were computed, with standard errors estimated using the estimator in Theorem 1 (Table 2, columns 2−3). The estimated increase in disability as a function of year is markedly weaker in this model fit. In addition, it appears as if the standard method ignoring drop-out over-estimates the effects of age, education and memory by between 14 and 32%, while the effect of sex is underestimated by about 20%. The fact that slopes with respect to time would be over-estimated in a standard analysis makes sense because subjects experiencing an increase in disability are more likely to drop-out soon thereafter, so that subsequent declines would not be detected.

Finally, we used a transition model to implement the improved efficiency estimator U^{appr} from Section 3.4. The transition model for disabilitiy at time t included year, age, sex, memory, disability at time $t-1$, as well as all two-way interactions involving year and disability (Table 2, columns 4−5). The estimates are generally closer to the bias-corrected estimates than to the estimates from standard CLR, as expected, and the standard errors are slightly smaller. Here, the efficiency improvements of U^{appr} over U^c are only on the order of 5%. This could be due to the richness of the drop-out model, which contained several interaction terms. From our analysis, we conclude that higher memory leads to slower declines in disability. The effect of age is as expected, while that of education is in the opposite direction. Additional analyses suggested that the observed education effect was due to regression to the mean as, cross-sectionally at baseline, eduction was negatively associated with disability.

6. Discussion

Under MAR, a full integrated likelihood-based estimator assuming a parametric random effect distribution is consistent, but the standard conditional likelihood estimator is biased. It may be puzzlling that a MAR process is not ignorable here, as ignorability is usually assumed to hold for likelihood-based estimators under MAR processes. The reason for this has to do with the likelihood used for conditioning, and can be seen as follows. First, suppose that q is fixed and that the observed dropout time T is t. Then, ignoring dependencies on $(\mathbf{X}, Z)$, the likelihood is

$$\Pr(T=t, \mathbf{Y}_{\text{obs}}) = \Pr(\mathbf{Y}_t|T=t)\Pr(T=t) = \Pr(T=t|\mathbf{Y}_t)\Pr(\mathbf{Y}_t).$$

In this last form for the likelihood, if q is not to be eliminated from the problem, then, since neither q nor β appear in $\Pr(T=t|\mathbf{Y}_t)$, this factor can be dropped

and likelihood $\Pr(\mathbf{Y}_t) = \Pr(\mathbf{Y}_{\text{obs}})$ can be used for inferences on (β, q). Similarly, if q is treated as a random variable, then the integrated likelihood is

$$\int_q \Pr(T = t, \mathbf{Y}_{\text{obs}})\, \mathrm{d}q = \Pr(T = t|\mathbf{Y}_t) \int_q \Pr(\mathbf{Y}_t)\, \mathrm{d}q,$$

where again $\Pr(T = t|\mathbf{Y}_t)$ does not depend on q. Again, inferences can be based on $\int_q \Pr(\mathbf{Y}_{\text{obs}})\, \mathrm{d}q$, ignoring the factor $\Pr(T = t|\mathbf{Y}_t)$. However, when q is to be eliminated from the problem using conditioning, one must do the conditioning on a proper probability mass function. In this paper we work with $\Pr(\mathbf{Y}_{\text{obs}}|T)$, for which further conditioning on $\sum_{t=1}^{T} Y_t$ eliminates q. By contrast, $\Pr(\mathbf{Y}_{\text{obs}}) = \Pr(\mathbf{Y}_T)$ by itself is not a proper probability mass function because it does not account for the role of T as either a random variable or a conditioning statistic. Valid probability functions are $\Pr(\mathbf{Y}_T|T)$, which we use in this paper, and $\Pr(\mathbf{Y}_T, T)$. Because our starting point is $\Pr(\mathbf{Y}_T|T)$, the drop-out process plays a role. We do not use the joint distribution $\Pr(\mathbf{Y}_T, T)$ because without further conditioning on T, it is very difficult to eliminate q from the problem.

We have assumed that the covariate vector X_{it} for subject i is observable even for $t > T_i$. This assumption will hold if X_{it} is a function of baseline covariates Z_i and time t and/or if X_{it} is measured through an external process. More generally, we might consider a model in which X_{it} is replaced by (X_{it}, W_{it}), where W_{it} is a vector of covariates that are measured concurrently with Y_{it} and which cannot be measured for $t > T_i$. The method developed in Section 2 easily extends to this setting, as long as $\lambda_i(t, \mathbf{y}_{t-1})$ depends only on $\mathbf{W}_i = (W_{i1}, \ldots, W_{iJ})'$ through $(W_{i1}, \ldots, W_{i,t-1})'$. Then, to extend the projection method in Section 3.3 to incorporate W_{it}, we would require that the expected value in $V_{\text{proj}}^{(t,s)}$ to be taken over $(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{R}_i)$. The transition-model based approximation in Section 3.4 would still apply, providing that (14) were extended to a joint transition model for (Y_{it}, W_{it}).

Acknowledgement

This material is based upon work supported by the U.S. National Science Foundation, Grant No. 0096412, and by the U.S. National Institute on Aging, Grant No. R03 AG-18803-01. The author thanks Diane Lauderdale, Ph.D., for invaluable input on the aging example analysis, and an associate editor and the referees for diligent reviews that significantly improved the paper.

Appendix A. Sketch proof of Theorem 1

The argument is similar to that of Theorem 1 of Robins, Rotnitzky and Zhao (1995). Consistency of $\widehat{\beta}$ follows from standard pseudo-likelihood theory (Gong

and Samaniego (1981)). The asymptotic normality involves a standard Taylor series argument, wherein the form of the asymptotic variance $\mathcal{V}$ follows from the facts that both U^c and S^γ are likelihood scores, so that minus the expected value of their derivatives is equal to their variances, and that $\mathrm{E}\,(-\partial U^a/\partial\beta) = \mathrm{E}\,(-\partial S^\gamma/\partial\beta) = 0$.

References

Baker, S. G. (1995). Marginal regression for repeated binary data with outcome subject to non-ignorable non-response. *Biometrics* **51**, 1042-1052.

Begun, J. M., Hall, W. J., Huang, W.-M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist.* **11**, 432-452.

Bertschek, I. and Lechner, M. (1998). Convenient estimators for the panel probit model. *J. Econometrics* **87**, 327-371.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9-25.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research*, vol. 1, *The Analysis of Case-Control Studies.* International Agency for Research on Cancer, Lyon.

Conaway, M. R. (1992). The analysis of repeated categorical measurements subject to nonignorable nonresponse. *J. Amer. Statist. Assoc.* **87**, 817-824.

Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Appl. Statist.* **43**, 49-93

Diggle, P. J., Liang, K.-Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Oxford University Press, Oxford.

Fitzmaurice, G. M., Molenberghs, G. and Lipsitz, S. R. (1995). Regression models for longitudinal binary responses with informative drop-outs. *J. Royal Statist. Soc. Ser. B* **57**, 691-704.

Gong, G. and Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: theory and applications. *Ann. Statist.* **9**, 861-869.

Greene, W. H. (2003). *Econometric Analysis.* 5th edition. Prentice-Hall, Upper Saddle River, New Jersey.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *J. Amer. Statist. Assoc.* **90**, 1112-1121.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* John Wiley and Sons, New York.

Newey, W. K. (1990). Semiparametric efficiency bounds. *J. Appl. Econometrics* **5**, 99-135.

Rathouz, P. J. (2003). Likelihood methods for missing covariate data in highly stratified studies. *J. Royal Statist. Soc. Ser. B* **65**, 711-723.

Rathouz, P. J., Satten, G. A. and Carroll, R. J. (2002). Semiparametric inference in matched case-control studies with missing covariate data. *Biometrika* **89**, 905–916.

Robins, J. M., Rotnitzky, A. and van der Laan, M. (2000). Comment on 'On profile likelihood,' by Murphy, S. A. and van der Vaart, A. W. *J. Amer. Statist. Assoc.* **95**, 477-482.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90**, 106-121.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* **90**, 122-129.

Soldo, B. J., Hurd, M. D., Rodgers, W. L. and Wallace, R. B. (1997). Asset and health dynamics of the oldest old: an overview of the AHEAD study. *J. Gerontology, Ser. B, Psychological Sci. and Social Sci.* **52B**, 1-20.

Ten Have, T. R., Kunselman, A. R., Pulkstenis, E. P. and Landis, J. R. (1998). Mixed effects logistic regression models for longitudinal binary response data with informative drop-out. *Biometrics* **54**, 367-383.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* **44**, 175-188.

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

Department of Health Studies, University of Chicago, 5841 South Maryland Avenue, MC 2007, Chicago, IL 60637, U.S.A.

E-mail: prathouz@health.bsd.uchicago.edu

(Received January 2003; accepted November 2003)