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 those 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 re-
gression; Missing data; Nuisance parameter; Projection; Double-robustness; Subject-
specific model.

Address for correspondence: Paul J. Rathouz, Department of Health Studies, 5841
South Maryland Avenue, MC 2007, Chicago, IL 60637.
Ph: 773.667.9905. FAX: 773.702.1979. E-mail: prathouz@health.bsd.uchicago.edu
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

$$\text{logit}\{E(Y_{it}|X_i)\} = q_i + X_{it}'\beta,$$

where $Y_{it}$ is a binary outcome variable on subject $i$ at time $t$, $X_{it}$ is a vector of time-varying covariates, $\beta$ is a vector of regression coefficients, and $X_i$ is the matrix $X_i = (X_{i1}, \ldots, X_{iJ})'$. In this logistic model, $q_i$ is a subject-specific intercept that accounts for the fact that the components of the vector $Y_i = (Y_{i1}, \ldots, Y_{iJ})'$ 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 $\beta$ 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; Diggle, Liang and Zeger, 1994, sect. 9.3.1). While both models yield a subject-specific interpretation of $\beta$, 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 $X_i$, although it is certainly possible to model the dependence of $q_i$ on $X_i$. By contrast, in fixed effects models, $q_i$ captures all subject level factors, including those related to $X_i$. 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 $i$th 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 $X_i$ in an unknown way. This yields a semiparametric model, with $X'_{it}\beta$ being the parametric part, and the distribution of $q_i$ given $X_i$ being nonparametric. The CLR estimator is semiparametric efficient for $\beta$ 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 (Lauderdale et al, submitted). The data used to address this question are from the Study of Assets and Health Dynamics Among the Oldest Old (AHEAD; Soldo et al, 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 § 5.

Attrition is an important problem in aging research such as this. For example, our
data set comprises 6350 subjects, but only 3705 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 $Y_i$ and $X_i$, analysis based on the standard conditional likelihood generated from the first $T_i$ observations will yield consistent and asymptotically normal estimators for $\beta$. However, if drop-out depends on $Y_i$ and/or $X_i$, a standard complete record analysis 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

$$I(T_i = t) \Pi Y_{i,t+1}, \ldots, Y_{iJ} | Y_{i1}, \ldots, Y_{it}, X_i, Z_i,$$  \hspace{1cm} (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 $Y_{i1}, \ldots, Y_{it}, X_i, Z_i$.

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

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

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 the discussion (§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 §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 $X_{it}$, or on random effects models. The marginal model approaches either rely on a model for the drop-out process to account for missingness in a generalized estimating equations framework (Robins, Rotnitzky and Zhao, 1995; Robins and Rotnitzky, 1995; Fitzmaurice, Molenberghs and Lipsitz, 1995) or take a likelihood approach, where the likelihood is expressed in terms of marginal model parameters (e.g., Baker, 1995; Diggle and Kenward, 1994). In the random effects setting, drop-out is often tied to the response data through the latent random coefficients (e.g., Wu and Carroll, 1988; Ten Have, Kunselman, Pulkstenis, and Landis, 1998) in what are sometimes called ‘shared-parameter’ models. 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 here 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 § 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 $\beta$, primarily because it relies on information in the observed drop-out process, which is ancillary for $\beta$. In § 3, we address this problem by identifying a class of estimating functions of which the score function developed in § 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 $\beta$-estimation over the score in § 2. We also show that this efficient estimating function is doubly-robust (Robins, Rotnitzky and van der Laan, 2000). 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 § 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 § 2 takes a different form than that in our previous work. Moreover, the projection argument and subsequent approximation in §§ 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 assumption (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 \( i \)th subject’s ‘drop-out time’. Let \( R_{it} = 1 \) or \( 0 \) indicate whether the \( t \)th observation \((Y_{it}, X_{it})\) is observed or missing for the \( i \)th subject. That is, that \( R_{it} = 1 \) iff \( t \leq T_i \).

To denote the data for a given subject \( i \), write \( Y_i = (Y_{i1}, \ldots, Y_{iJ})' \) for the \( J \times 1 \) vector of binary responses, including any values missing due to drop-out. Similarly
define the $J \times p$ matrix $X_i = (X_{i1}, \ldots, X_{ij})'$ of $p \times 1$ row vectors of covariates for subject $i$, and the vector $R_i$ of missing data indicators. Further define $Y_{i,\text{obs}}$ to be the $T_i$ observed components of $Y_i$, and $X_{i,\text{obs}}$ to be the $T_i$ observed rows of $X_i$, before or at drop-out. Note that which components of $Y_i$ and $X_i$ appear in $Y_{i,\text{obs}}$ and $X_{i,\text{obs}}$ is a function of the indicator vector $R_i$. Finally, a subscript $t$ added to a $Y_i$ or $X_i$ 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, $Y_{it} = (Y_{it1}, \ldots, Y_{itJ})'$. Of course, with some redundancy, we then have $Y_i \equiv Y_{iJ}$ and $Y_{i,\text{obs}} \equiv Y_{iT_i}$.

We express model of interest (1) in terms of the odds that $Y_{it} = 1$. As such, let $\theta_{it} = \theta_i(X_{it}; \beta) = \Pr(Y_{it} = 1|X_i)/\Pr(Y_{it} = 0|X_i)$. The goal is to make inferences about the $p \times 1$ parameter $\beta$ 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

$$
\lambda_i(t, Y_{i,t-1}; \gamma) = \Pr(R_{it} = 1|R_i = \ldots = R_{i,t-1} = 1, Y_i = y, X_i, Z_i)
$$

$$
= \Pr(R_{it} = 1|R_{i,t-1} = 1, Y_i = y, X_i, Z_i),
$$

where $\gamma$ is a finite-dimensional nuisance parameter which does not depend on response model parameters $(q_i, \beta)$, $y = (y_1, \ldots, y_J)'$ and the role of subscript $t - 1$ on the bold $y$ is as defined above. Note that $1 - \lambda_i(t, Y_{i,t-1})$ is the hazard of drop-out at $t$, and that MAR (2) ensures that $\lambda_i(t, Y_{i,t-1})$ does not depend on $Y_{it}, \ldots, Y_{iJ}$. For example, $\lambda_i(t, Y_{i,t-1})$ 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, $\lambda_i(t, Y_{i,t-1})$ might vary across $t$ and/or depend on more than just the most recent
value of \( Y_{it'}, t < t' \). Also, \( \lambda_i(t, Y_{i,t-1}) \) is implicitly allowed to depend on subject-level covariates \( Z_i \), such as sex or treatment assignment, as well as the matrix \( X_i \) of time-varying covariates, which might include time or treatment-by-time interactions. Throughout, we assume that dependence of \( \lambda_i(t, y_{t-1}) \) on \( (X_i, Z_i) \) is indicated by subscript \( i \), but we make the dependence on \( y_{t-1} \) explicit for reasons that will become clear. We also assume that \( \lambda_i(1, y_0) \equiv 1 \), i.e., the baseline assessment is always observed, and, for ease of exposition, that \( \lambda_i(J + 1, y_J) \equiv 0 \).

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

\[
\pi_i(t, y_t; \gamma) = \Pr(T_i = t|Y_i = y, X_i, Z_i) = \prod_{s=1}^{t} \lambda_i(s, y_{s-1}; \gamma) \{1 - \lambda_i(t + 1, y_t; \gamma)\}.
\]

By MAR assumption (2), \( \pi_i(t, Y_{it}) \) only depends 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 \( Y_{i1}, \ldots, Y_{iT_i}, X_i, Z_i \). This likelihood, \( L_i \), is conditional on the drop-out process \( T_i \), that is

\[
L_i(\beta, \gamma, q_i) = \Pr(Y_i, \text{obs} | X_i, Z_i, T_i).
\]

Expressing this likelihood via odds \( \theta_{it} \) and drop-out probability \( \pi_i(t, y_t) \), yields

\[
L_i(\beta, \gamma, q_i) = \sum_{Y_{*i,\text{obs}}} \prod_{t=1}^{T_i} \theta_{it}^{Y_{it}} \prod_{t=1}^{T_i} \pi_i(T_i, y_{\text{obs}}) \frac{Y_{*i,\text{obs}}}{\sum_{Y_{*i,\text{obs}}} \prod_{t=1}^{T_i} \theta_{it}^{Y_{it}} \pi_i(T_i, y_{\text{obs}})}, \tag{3}
\]

where \( Y_{*i,\text{obs}} \) is the set of all \( 2^{T_i} \) possible vectors \( y_{\text{obs}} = (y_1, \ldots, y_{T_i})' \).

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

\[
L_i'(\beta, \gamma) = \Pr(Y_{i,\text{obs}} | \sum_{t=1}^{T_i} Y_{it}, X_i, Z_i, T_i). \tag{4}
\]
Then
\[ L_i^c(\beta, \gamma) = \frac{\{\prod_{t=1}^{T_i}(e^{X_{i,t}^\beta} y_{it})\} \pi_i(T_i, Y_{i,\text{obs}})}{\sum_{y_{obs} \in Y_{i,\text{obs}} \{\prod_{t=1}^{T_i}(e^{X_{i,t}^\beta} y_{it})\} \pi_i(T_i, y_{obs})}, \]
where \( Y_{i,\text{obs}} = Y_{i,\text{obs}}(Y_{i,\text{obs}}) \) is the set of all vectors \( y_{obs} = (y_1, \ldots, y_{T_i})' \) 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 \( \pi_i(T_i, y_{obs}) \) is constant across all \( y_{obs} \in Y_{i,\text{obs}} \). 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 \( \gamma \), we model \( R_{i,t} \) among those subjects for whom \( R_{i,t-1} = 1 \). As such, define \( L_i^\gamma \) to be the \( i \)th subject’s contribution to the \( \gamma \)-likelihood, i.e.,
\[ L_i^\gamma(\gamma) = \left\{ \prod_{t=1}^{T_i-1} \lambda_i(t, Y_{i,t-1}; \gamma) \right\} \{1 - \lambda_i(T_i, Y_{i,T_i-1}; \gamma)\}. \]
Then, accumulating information over subjects \( i = 1, \ldots, K \), let \( \hat{\gamma} \) be the estimator of \( \gamma \) obtained by maximizing the likelihood \( \prod_i L_i^\gamma(\gamma) \).

We propose maximum conditional likelihood estimation of \( \beta \) using \( L_i^c \), with \( \hat{\gamma} \) replacing \( \gamma \). Combining information across subjects, let \( \hat{\beta} \) be the maximizer of \( \prod_i L_i^c(\beta, \hat{\gamma}) \).

Equivalently, \( \hat{\beta} \) is the solution to \( \sum_i U_i^c(\beta, \hat{\gamma}) = 0 \), where \( U_i^c(\beta, \gamma) \) is the \( i \)th subject’s \( \beta \)-score contribution \( U_i^\beta = (\partial \log L_i^c/\partial \beta) \), \( \hat{\gamma} \) is the solution to \( \sum_i S_i^\gamma(\gamma) = 0 \), and \( S_i^\gamma(\gamma) \) is the \( i \)th subject’s \( \gamma \)-score contribution \( S_i^\gamma(\gamma) = (\partial \log L_i^\gamma/\partial \gamma) \). An estimator for the asymptotic variance of \( \hat{\beta} \) with estimated \( \gamma \) is given as a special case of Theorem 1 in § 3.2 by replacing \( U_i \) in that theorem with \( U_i^c \).

**Remark.** In the special case where \( \lambda_i(t, Y_{i,t-1}) \) only depends on the most recent response value, that is, \( \lambda_i(t, y_{t-1}) = \lambda_i(t, y_{t-1}) \), \( \hat{\beta} \) can be computed using standard
software, as follows. First define
\[
B_{it} = \begin{cases} \log\{\lambda_i(t+1)/\lambda_i(t,0)\}, & t = 1, \ldots, T_i - 1 \\ \log\{1 - \lambda_i(t+1)\}/\{1 - \lambda_i(t,1)\}, & t = T_i \end{cases}
\]
We show in Appendix A that
\[
L_i^c(\beta, \gamma) = \frac{\prod_{t=1}^{T_i}(e^{X_{it}'\beta} + B_{it})y_{it}}{\sum_{y_{obs}\in Y_{i,obs}} \prod_{t=1}^{T_i}(e^{X_{it}'(\beta + B_{it})}y_{it})},
\]
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 \(\beta\) will be consistent, although the standard errors produced by the package will be conservative.

3. Efficiency Improvements

3.1. Introduction

While conditional likelihood \(L_i^c\) will yield valid inferences about \(\beta\), it is potentially inefficient because it contains information on the missingness process \(R_i\), which is ancillary for \(\beta\). To see this, note that likelihood \(L_i^c\) depends in no way on \(\beta\), and yet \(L_i^c\) depends on random variables \(R_i\). This suggests that more efficient estimation of \(\beta\) can be achieved by using an estimating function where the \(\beta\)-ancillary information in \(R_i\) has been removed. Heuristically, the idea is to identify an estimating function \(U_i^a\) that (i) contains no \(\beta\)-information, i.e., is ancillary for \(\beta\), (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 \(\beta\)’ to mean that \(E(-\partial U_i^a/\partial \beta) = 0\). Then, \(U_i^c - U_i^a\) will yield a potential increase in efficiency relative to \(U_i^c\).

In § 3.2, equations (7) and (9), we introduce a class \(U_i\) of estimating functions for \(\beta\) that includes the complete-record estimating function \(U_i^c\) introduced in § 2.3 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 § 3.3, where we use a projection
argument and semiparametric efficiency theory to generate the optimal member $U_i^{\text{proj}}$ of $\mathcal{U}$. We show that $U_i^{\text{proj}}$ satisfies (iii) and yields more efficient $\beta$ inferences than $U_i^c$. Finally, in § 3.4, we propose an approximation $U_i^{\text{appr}}$ to $U_i^{\text{proj}}$, also in $\mathcal{U}$, that is more practical to compute than $U_i^{\text{proj}}$ and that is also expected to satisfy (iii). Simulations in § 4 illustrate the potential efficiency improvements in $U_i^{\text{appr}}$ over $U_i^c$.

Throughout § 3, we treat $q_i$ as an unobserved random variable with a nonparametric distribution depending arbitrarily on $(X_i, Z_i)$. We note that conditional likelihood $L_i^c$ and scores $U_i^c$ remain perfectly valid under this model. References to the joint distribution of $Y_i$ or of its sub-vectors $Y_{it}$ are conditional on $(X_i, Z_i)$, but marginal over $q_i$. That is, $(Y_i|X, Z_i)$ has a nonparametric mixture distribution. Throughout § 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 $V^{(t, s)}(R_s, Y_{s-1}, X, Z; \beta, \gamma, \alpha)$. Generally, $V^{(t, s)}$ will depend on $\beta$, $\gamma$ and possibly a finite dimensional nuisance parameter $\alpha$. Set $V^{(J, J+1)} \equiv 0$.

Now, consider estimating functions of the form

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

(7)

where we define $R_0 \equiv 1$ and, taking expectation over $R_s$,

$$\epsilon^{(t, s)} = E(V^{(t, s)}|R_{s-1} = 1, Y_{s-1}, X, Z).$$

Note that, owing to the factor $R_{s-1}$ in (7), $U^a$ can always be computed with observed data. We show in Appendix B that, regardless of choice of $V^{(t, s)}$,

$$E(U^a) = E(-\partial U^a/\partial \beta) = 0,$$

(8)
as long as the missingness model \( \lambda_i(t, y_{t-1}) \) is correctly-specified, so that (7) defines a class of unbiased \( \beta \)-ancillary estimating functions \( U^\alpha \), satisfying criteria (i) and (ii) in \S 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 \( \beta \) of the form

\[
U(\beta, \gamma, \alpha) = U^c(\beta, \gamma) - U^a(\beta, \gamma, \alpha).
\]  

(9)

Specific members of \( \mathcal{U} \) are the subjects of \S\S 3.3-3.4.

For a given choice of \( V^{(t,s)} \), to use the resulting estimating function \( U \in \mathcal{U} \) for \( \beta \)-inferences, assume that there exists an \( \alpha \)-estimating function \( S^\alpha = S^\alpha(Y_{obs}, X_{obs}, Z, T; \alpha) \), and let \( \tilde{\alpha} \) be the solution to \( \sum_i S^\alpha_i(\alpha) = 0 \). Suppose that we estimate \( \gamma \) as in \S 2.3, and \( \beta \) by solving \( \sum_i U_i(\beta, \tilde{\gamma}, \tilde{\alpha}) = 0 \). Theorem 1 characterizes the asymptotic distribution of \( \tilde{\beta} \). Proof is in Appendix C.

**Theorem 1.** Suppose that the modeling assumptions in \S 2 hold, that \( \tilde{\gamma} \) solving \( \sum_i S^\alpha_i(\gamma) = 0 \) is \( \sqrt{K} \)-consistent, that \( \tilde{\alpha} \) as defined above is \( \sqrt{K} \)-consistent for some \( \alpha^* \), and that \( \tilde{\beta} \) solves \( \sum_i U_i(\beta, \tilde{\gamma}, \tilde{\alpha}) = 0 \). Then, under mild regularity conditions as \( K \to \infty \), \( \tilde{\beta} \to \beta \) in probability, and \( \sqrt{K}(\tilde{\beta} - \beta) \to N(0, \mathcal{V}) \) in distribution, with

\[
\mathcal{V} = \lim_{K \to \infty} \mathcal{V}_K,
\]

where

\[
\mathcal{V}_K = K(\sum_i T^c_{\beta i})^{-1}(\sum_i \tilde{U}_i \tilde{U}_i^T)(\sum_i T^a_{\beta i})^{-1},
\]

(10)

\( T^c_{\beta i} = E(U^c_i U^{c^T}_i) \), and \( \tilde{U}_i \) is the residual from the least-squares regression of the \( U_i \)'s on to \( \gamma \) scores \( S^\gamma_i \); that is, \( \tilde{U}_i = U_i - CS^\gamma_i \), where \( C = (\sum_i U_i S^\gamma_i^T)(\sum_i S^\gamma_i S^\gamma_i^T)^{-1} \).

**Remark.** From (10), we see that the asymptotic efficiency of \( \tilde{\beta} \) for \( U^c \) or any \( U \in \mathcal{U} \) is improved by estimation of \( \gamma \), even if \( \gamma \) is already known. Similarly, using a model for the drop-out process \( \lambda(t, y_{t-1}) \) 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 § 4.

3.3. Improved efficiency via projection

We now exploit ideas in semiparametric efficiency theory to obtain a member \( U^{\text{proj}} \) that is optimal in the class \( \mathcal{U} \) defined in § 3.2. The key idea is to remove from \( U^c \) its projection onto the tangent space \( \mathcal{W} \) for the nuisance parameter \( \gamma \) (i.e., the closed linear span of \( \mathcal{L}^2 \) scores for \( \gamma \); Newey, 1990, § 3; Robins, Rotnitzky and van der Laan, 2000, § 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 \((R_s, R_{s-1}, Y_{s-1}, X, Z)\) which are unbiased conditional on \((R_{s-1}, Y_{s-1}, X, Z)\). 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 \( g \),

\[
\mathcal{P}_s g = E(g|R_s, R_{s-1}, Y_{s-1}, X, Z) - E(g|R_{s-1}, Y_{s-1}, X, Z).
\]

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 \( R \)) 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), \( L^{(t)}(\beta, \gamma) = \Pr(Y_t|\sum_{s=1}^{t} Y_{s}, X, Z, T = t) \). Then, \( L^{c} = \prod_{t=1}^{J} L^{(t)} I(T = t) \). Similarly, writing \( U^{(t)} = (\partial \log L^{(t)}/\partial \beta) \), we can rewrite \( U^c \) as

\[
U^c = \sum_{t=1}^{J} I(T = t) U^{(t)}. \tag{11}
\]

To compute projections \( \mathcal{P}_s U^c \), we exploit decomposition (11), operating one term
at a time. We show in Appendix D that, for $1 \leq s \leq t + 1$, $1 \leq t \leq J$,

$$\mathcal{P}_s I(T = t) U(t) = R_{s-1} \left( V_{\text{proj}}^{(t,s)} - \epsilon_{\text{proj}}^{(t,s)} \right), \quad (12)$$

where

$$V_{\text{proj}}^{(t,s)}(\beta, \gamma, \alpha) = E \left\{ (1 - R_{t+1}) R_t U(t) | R_s, R_{s-1} = 1, Y_{s-1}, X, Z \right\},$$

expectation being taken over $(Y, R)$, and

$$\epsilon_{\text{proj}}^{(t,s)}(\beta, \gamma, \alpha) = E \left( V_{\text{proj}}^{(t,s)} | R_{s-1} = 1, Y_{s-1}, X, Z \right),$$

expectation being over $R_s$. We also show that, for $s > t + 1$, $\mathcal{P}_s I(T = t) U(t) = 0$ and, for ease of exposition, we define $\mathcal{P}_{J+1} g \equiv 0$. Further technical details on the form of $\mathcal{P}_s I(T = t) U(t)$ are given in Appendix D.

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

$$\mathcal{P} U^c = \sum_{t=1}^{J} \sum_{s=1}^{t+1} R_{s-1} \left( V_{\text{proj}}^{(t,s)} - \epsilon_{\text{proj}}^{(t,s)} \right), \quad (13)$$

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 $\beta$. Since $V_{\text{proj}}^{(t,s)}$ is a function of $(R_s, Y_{s-1}, X, Z)$, $\mathcal{P} U^c$ satisfies (7), and $U^{\text{proj}}$ is therefore in the class $\mathcal{U}$ defined by (9). Theorem 2 elucidates the efficiency benefit in using $U^{\text{proj}}$ over $U^c$, or any other $U \in \mathcal{U}$, for inferences about $\beta$, and implies that projection (13) satisfies criterion (iii) in § 3.1. In addition, $U^{\text{proj}}$ is a doubly robust estimating function for $\beta$ in the sense that, for any $(\gamma^\dagger, \alpha^\dagger)$, if either $\gamma^\dagger = \gamma$ or $\alpha^\dagger = \alpha$, then $U^{\text{proj}}(\beta, \gamma^\dagger, \alpha^\dagger)$ is unbiased (Robins, Rotnitzky and van der Laan, 2000, Lemma 1). Sketch proofs of double robustness and Theorem 2 are in Appendix E.

**Theorem 2.** $U^{\text{proj}}$ 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 $U^{\text{proj}}$ information matrix $\mathcal{I}_\beta^{\text{proj}}$ is $\mathcal{I}_\beta^{\text{proj}} = \mathcal{I}_\beta E(U^{\text{proj}} U^{\text{proj}}^T)^{-1} \mathcal{I}_\beta^c$ and $\mathcal{I}_\beta$ is the corresponding information matrix for $U$. 

15
Remarks.

1. The proof of Theorem 2 involves three main points, which hold for any $U \in U$. First, $E(-\partial U/\partial \beta) = I_\beta$. Second, $U - PU = U^c - PU^c = U^{proj}$. Third, $PU$ is positively-correlated with $U$, so that

$$E(UU^T) - E(U^{proj}U^{proj T}) \geq 0$$

in the positive semidefinite sense.

2. Theorem 1 demonstrated, for a given choice $U \in U$, that efficiency is improved by estimation of $\gamma$ 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.

3. Note that $PU^c$, and hence $U^{proj}$, depend on the joint distribution of $(Y|X, Z)$ to compute $V^{(t,s)}_{proj}$, and the distribution of $(Y|X, Z)$ in turn depends on a model for the mixture distribution of $(q|X, Z)$. This is the first place in our development where $(q|X, Z)$ is needed. However, note that if the model for $(Y|X, Z)$ is misspecified, the resulting $V^{(t,s)}_{proj}$’s will not be the correct, optimal functions, but they will still yield valid functions $V^{(t,s)}$ as defined in § 3.2. Also, given any $V^{(t,s)}_{proj}$, correct computation of $e^{(t,s)}$ ($e^{(t,s)}_{proj}$) only depends on the drop-out model, $\lambda(t; y_{t-1})$. The implication is that even if the joint distribution of $(Y|X, Z)$ is misspecified, so long as the drop-out model is correct, the resulting $PU^c$, while not optimal, will still be a valid member of the class $(7)$, and the resulting $U^{proj} = U^c - PU^c$ will still be a valid unbiased estimating function in the class $U$.

4. Even if the joint distribution $(Y|X, Z)$ and resulting $PU^c$ 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 $PU^c$ will be positively correlated with $U^c$ (criteria (iii)), so that (14) in remark 1. will still hold.
3.4. A practical estimator

The results of the foregoing section show that, to compute the $V_{\text{proj}}^{(t,s)}$'s and the projection $\mathcal{P}U^c$ needed for $U^{\text{proj}}$, we require the full joint distribution $(\mathbf{Y}|\mathbf{X}, Z)$, marginally over $q$. In computing $U^{\text{proj}}$, distribution $(\mathbf{Y}|\mathbf{X}, Z)$ plays a role in increasing efficiency of $\hat{\beta}$, but correct specification of this distribution is not critical for consistency. Also, it may be difficult to compute the $V_{\text{proj}}^{(t,s)}$'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 $(\mathbf{Y}_t|\mathbf{Y}_{t-1}, \mathbf{X}, Z)$. 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 $V_{\text{app}}^{(t,s)}$ to $V_{\text{proj}}^{(t,s)}$. $\epsilon_{\text{app}}^{(t,s)}$ is then obtained from $V_{\text{app}}^{(t,s)}$ as in (7).

Define the transition probabilities

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

for $t = 2, \ldots, J$, where $\alpha$ is a finite-dimensional nuisance parameter. In principle, model $\eta(\cdot; \alpha)$ and parameter $\alpha$ depend on the interest parameter $\beta$. 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, working model (15) meets our needs for a practical approximation to distribution $(\mathbf{Y}|\mathbf{X}, Z)$.

Model (15) for $(\mathbf{Y}|\mathbf{X}, Z)$, together with the model $\lambda(t, y_{t-1}; \gamma)$ for the drop-out 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 $P \mathcal{U}^c$. As such, define functions $V_{\text{appr}}^{(t,s)}$ as

$$V_{\text{appr}}^{(t,s)}(\gamma, \alpha) = \tilde{E}\left\{(1 - R_{t+1})R_t U^{(t)} R_s, R_{s-1} = 1, Y_{s-1}, X, Z; \alpha, \gamma\right\},$$

where $\tilde{E}(:, \alpha, \gamma)$ denotes expectation taken with respect to the working distribution of $(R, Y|X, Z)$. Note importantly that functions $V_{\text{appr}}^{(t,s)}$ are of the form $V^{(t,s)}$ given in § 3.2, and hence they define an element $U_{\text{appr}}$ 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 $U_{\text{appr}}$ as an improvement over $U^c$ for inferences on $\beta$.

In order to use $U_{\text{appr}}$, we require an estimator of transition model parameter $\alpha$. For this, note that, under MAR assumption (2), the $i$th subject’s contribution to the likelihood function $L_i^{\alpha}$ for $\alpha$ is given by

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

Accumulating information over subjects $i = 1, \ldots, K$, let $\hat{\alpha}$ be the estimator of $\alpha$ obtained by maximizing the likelihood $\prod_i L_i^{\alpha}(\alpha)$. We assume that $\hat{\alpha}$ is consistent for some value $\alpha^*$ even if transition model (15) is not valid or is misspecified. Note that, similarly to $\hat{\gamma}$, $\hat{\alpha}$ can be computed via $L_i^{\alpha}$ before estimation of $\beta$. With estimators $\hat{\gamma}$ and $\hat{\alpha}$, $U_{\text{appr}}^{(\beta, \hat{\gamma}, \hat{\alpha})}$ is an approximate projected estimating function, and the resulting estimator $\hat{\beta}$ solving $\sum_{i} U_{i}^{\text{appr}}(\beta, \hat{\gamma}, \hat{\alpha}) = 0$ has asymptotic distribution given in Theorem 1. We investigate the performance of $U_{\text{appr}}$, and compare it to $U^c$, via simulations in the next section.

4. Simulation study

To investigate the performance of the estimators based on $U^c$ and $U_{\text{appr}}$, as compared with the naive complete-record estimator, 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 logit\{Pr($Y_{it} = 1|X_i, q_i$)} = $q_i + \beta_t(t - 1)/4 + \beta_x X_{it}$, 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

\[
\logit\{\Pr(R_{it} = 1| R_{i,t-1} = 1, Y_{i,t-1}, X_i, Z_i)\} = \gamma_0 + \gamma_t(t - 1)/4 + \gamma_y Y_{i,t-1} + \gamma_z Z_i \quad (16)
\]

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 naive estimator computed using standard conditional logistic regression applied to the $T_i$ observed records for each subject $i$. The next three estimators were based on the bias-corrected likelihood $L_c^i$ and score $U_c^i$ proposed in § 2.3. For the second estimator, we assumed that $\lambda_i(t; y_{t-1})$, and hence $\pi_i(t, y_{t-1})$, were known. In the third, we estimated $\lambda_i(t; y_{t-1})$ using drop-out model (16), performing maximum likelihood estimation with $L_i^\gamma$ to obtain $\hat{\gamma}$, as described in § 2.3. We refer to (16) 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 (16). We call this the ‘rich’ drop-out model. We include it to evaluate potential increases in efficiency due to over-specification the missingness model. Finally, three estimators based on $U^{\text{appr}}$ were computed. The first used a ‘minimal’ transition model, modeling $Y_{it}$ as a function of $Y_{i,t-1}$ only. Here, the ‘minimal’ drop-out model (16) was used. In the second and third, a ‘full’ transition model was used, modeling

\[
\logit\{\Pr(Y_{it} = 1| Y_{i,t-1}, X_i, Z_i, q_i)\} = \gamma_0 + \alpha_t(t - 1)/4 + \alpha_y Y_{i,t-1} + \alpha_z Z_i \quad \text{for } t = 2, \ldots, 5.
\]
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 (10).

Results based on 1000 replicates are reported in Table 1. These include percent bias, mean square error efficiency relative to the bias-corrected estimator with estimated $\gamma$ under the ‘minimal’ drop-out model, and confidence interval coverage probabilities. As can be seen, naive estimation in this setting yields strongly biased estimators. All of the new estimators correct this bias. When using the bias-corrected $U^c$, estimation of $\gamma$ improves efficiency in $\hat{\beta}$ relative to the case where $\gamma$ 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 $PU^c$ has effectively already accounted for all such improvements.

5. Disability–Memory Example

We now revisit the aging research example introduced in § 1. In that study, baseline (year 0) data were collected in 1993, and follow up data were collected 2, 5 and 7 years later. Disability here is 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 10 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.

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 the interaction between education and disability. Other two-way interaction terms contributed very little to the fit of the dropout model. Using this drop-out model with \( U^c \) and the method in § 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 naive method 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 the naive method
makes sense because subjects experiencing an increase in disability are more likely to drop-out soon thereafter, so that a subsequent declines would not be detected.

Finally, we used a transition model to implement the improved efficiency estimator $U_{\text{appr}}$ from § 3.4. The transition model for disability 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 bias-corrected estimates than to the naive estimates, as expected, and the standard errors are slightly smaller. Here, the efficiency improvements of $U_{\text{appr}}$ 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, education was negatively associated with disability.

6. Discussion

We have developed a method for estimation of the logistic fixed effects model for longitudinal binary data subject to drop-out that is missing at random. We showed that the standard conditional likelihood method in this context is biased, and we developed a modified conditional likelihood that corrects this bias by taking the drop-out process into account. Then, recognizing that efficiency could be threatened by the fact that the drop-out process is ancillary in the full likelihood, we developed a projection argument to improve efficiency of the bias-corrected estimator, and demonstrated the degree of improvement in one simulated setting.

Our method can be contrasted with the random effects approach wherein the
subject-level intercepts $q$ are modelled as random quantities, and inferences are carried out via an integrated likelihood that is marginal over $q$. Two advantages to the parametric random effects approach are, first, that subject-level covariate effects can be estimated. As our approach absorbs all subject-level effects into $q$, only within-subject effects are estimated in our method. The second advantage of the random effects approach is that no specification of the drop-out process is required. By contrast, owing to the conditioning argument in the fixed effects model estimation, the drop-out process must be incorporated into the conditional likelihood. The key advantage of the fixed effects approach developed here is that it is completely robust to specification of the random effects structure, the trade-off being that it is less robust to specification of the drop-out process.

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 puzzling 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 $(X, Z)$, the likelihood is

$$\Pr(T = t, Y_{\text{obs}}) = \Pr(Y_t | T = t) \Pr(T = t) = \Pr(T = t | Y_t) \Pr(Y_t).$$

In this last form for the likelihood, if $q$ is not to be eliminated from the problem, then, since neither $q$ nor $\beta$ appear in $\Pr(T = t | Y_t)$, this factor can be dropped, and likelihood $\Pr(Y_t) = \Pr(Y_{\text{obs}})$ can be used for inferences on $(\beta, q)$. Similarly, if $q$ is treated as a random variable, then the integrated likelihood is

$$\int_q \Pr(T = t, Y_{\text{obs}}) dq = \Pr(T = t | Y_t) \int_q \Pr(Y_t) dq,$$

where again $\Pr(T = t | Y_t)$ does not depend on $q$. Again, inferences can be based
on $\int q \Pr(Y_{\text{obs}}) dq$, ignoring the factor $\Pr(T = t|Y_t)$. 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 $\Pr(Y_{\text{obs}}|T)$, for which further conditioning on $\sum_{t=1}^{T} Y_t$ eliminates $q$. By contrast, $\Pr(Y_{\text{obs}}) = \Pr(Y_T)$ 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 $\Pr(Y_T|T)$, which we use in this paper, and $\Pr(Y_T, T)$. Because our starting point is $\Pr(Y_T|T)$, the drop-out process plays a role. We do not use the joint distribution $\Pr(Y_T, T)$ 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 § 2 easily extends to this setting, as long as $\lambda_i(t, y_{t-1})$ only depends on $W_i = (W_{i1}, \ldots, W_{ij})'$ through $(W_{i1}, \ldots, W_{i,t-1})'$. Then, to extend the projection method in § 3.3 to incorporate $W_{it}$, we would require that the expected value in $V_{\text{proj}}^{(t,s)}$ to be taken over $(Y_i, W_i, R_i)$. The transition-model based approximation in § 3.4 would still apply, providing that model (15) were extended to a joint transition model for $(Y_{it}, W_{it})$.

**Acknowledgment**

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 the Associate Editor and referees for diligent reviews that significantly improved the paper.

**Appendix A. Proof of (6)**

First note that

$$\pi_i(t, y_t) = \frac{\prod_{s=1}^{t-1} \lambda_i(s+1, y_s)}{\lambda_i(s+1, 0)} \left( \frac{1 - \lambda_i(t+1, y_t)}{1 - \lambda_i(t+1, 0)} \right)^{y_t}.$$  

Therefore, $L^c_i$ can be written as in (6) with

$$\theta_{it}^* = \theta_{it} \frac{\lambda_i(t+1, 1)}{1 - \lambda_i(t+1, 0)}, \quad t = 1, \ldots, T_i - 1$$  

$$\theta_{iT_i}^* = \theta_{iT_i} \frac{1 - \lambda_i(t+1, 1)}{1 - \lambda_i(t+1, 0)}$$

completing the proof.

**Appendix B. Proof of (8)**

Each term in (7) is 0, as is obvious from the definition of $\epsilon(t,s)$. So $E(U^a) = 0$. For $E(-\partial U^a/\partial \beta) = 0$, note that, with dependence on $(X,Z)$ implicit,

$$\epsilon(t,s) = V^{(t,s)}(R_s = 1, Y_{S-1})\lambda_s(Y_{S-1}) + V^{(t,s)}(R_s = 0, Y_{S-1})\{1 - \lambda_s(Y_{S-1})\},$$

and $\lambda_s(Y_{S-1})$ does not depend on $\beta$. Therefore,

$$- \frac{\partial \epsilon(t,s)}{\partial \beta} = E \left( - \frac{\partial V^{(t,s)}}{\partial \beta} \bigg| R_{S-1} = 1, Y_{S-1} \right),$$

and the result follows the same argument as that leading to $E(U^a) = 0$.

**Appendix C. Proof of Theorem 1**

The consistency of $\hat{\beta}$ follows from standard pseudo-likelihood theory (Gong and Samaniego, 1981). For the asymptotic normality, define the following information quantities. Let $I_\beta = E(-\partial U/\partial \beta)$ and note that by (8), $E(-\partial U^a/\partial \beta) = 0$. Therefore

$$I_\beta = E(-\partial U^c/\partial \beta) = I^c_\beta.$$  

Because $U^c$ is a likelihood score, $I^c_\beta = E(U^cU^cT)$. Note
that $S^\gamma$ is ancillary for $\beta$, i.e., $E(-\partial S^\gamma/\partial \beta) = 0$. Let $I_\gamma = E(-\partial U/\partial \gamma)$ and $I_\gamma^\gamma = E(-\partial S^\gamma/\partial \gamma)$. Because $S^\gamma$ is a likelihood score for $\gamma$, $I_\gamma = E(U S^\gamma^T)$ and $I_\gamma^\gamma = E(S^\gamma S^\gamma^T)$. Finally, $E(-\partial U^c/\partial \alpha) = 0$ and $E(-\partial U^a/\partial \alpha) = 0$ by the same argument as that for $E(-\partial U^a/\partial \beta)$. Therefore $I_\alpha = E(-\partial U/\partial \alpha) = 0$.

Now, adding subscripts $i$ to index subjects and following the usual Taylor-series arguments, $(\hat{\gamma} - \gamma) = \left(\sum_i T_{\gamma,i}\right)^{-1} \sum_i S_i^\gamma + o_p(1/\sqrt{K}) = O_p(1/\sqrt{K})$, and similarly $(\hat{\alpha} - \alpha) = O_p(1/\sqrt{K})$. By Taylor series expansion,

$$\sum_i U_i(\beta, \hat{\gamma}, \hat{\alpha}) = \sum_i U_i + \{\sum_i (\partial U_i/\partial \gamma)\}(\hat{\gamma} - \gamma) + \{\sum_i (\partial U_i/\partial \alpha)\}(\hat{\alpha} - \alpha) + o_p(\sqrt{K})$$

$$= \sum_i U_i - \{\sum_i I_{\gamma,i}\}(\sum_i I_{\gamma,i})^{-1} \sum_i S_i^\gamma + o_p(1) + o_p(\sqrt{K})$$

$$= \sum_i \{U_i - \{\sum_{i'} I_{\gamma,i'}\}(\sum_{i'} T_{\gamma,i'})^{-1} S_i^\gamma\} + o_p(\sqrt{K}),$$

which then leads to the asymptotic normality of $\hat{\beta}$ with mean 0 and variance

$$K\{\sum_i \partial U_i(\beta, \hat{\gamma}, \hat{\alpha})/\partial \beta\}^{-1}\{\sum_i U_i(\beta, \hat{\gamma}, \hat{\alpha})^2\}\{\sum_i \partial U_i(\beta, \hat{\gamma}, \hat{\alpha})/\partial \beta\}^{-1}.$$ 

Finally, $-\sum_i \partial U_i(\beta, \hat{\gamma}, \hat{\alpha})/\partial \beta = -\sum_i \partial U_i/\partial \beta + o_p(K) = \sum_i I_{\beta,i} + o_p(K) = \sum_i I_{\beta,i} + o_p(K)$ and $\sum_i U_i(\beta, \hat{\gamma}, \hat{\alpha})^2 = \sum_i E(U_i U_i^T) - (\sum_i I_{\gamma,i})(\sum_i I_{\gamma,i})^{-1}(\sum_i I_{\gamma,i})^T + o_p(K)$.

These last two expressions yield (10) as a “sandwich estimator” of $\mathcal{V}$ (Huber, 1967).

**Appendix D. Technical details for projection $\mathcal{P}U^c$**

Here, we provide additional technical details for the projected estimating function $\mathcal{P}U^c$ developed in § 3.3. To do this, we first recast the model for $Y$ as a semi-parametric mixture model, replacing the nuisance parameter $q$ with a nuisance mixing distribution. Then, considering a non-parametric model for the missingness process $R$, we write the likelihood for the observed data ($Y_{\text{obs}}, T, X, Z$). Using this likelihood, we derive the nuisance tangent space $\mathcal{W}$ for the missingness process $R$, and show that it is the space spanned by the union of $\mathcal{W}_s$. Finally, using these results, we obtain expressions for $\mathcal{P}_s U^{(t)}$. 

26
Suppose that instead of treating $q$ as a nuisance parameter, we instead model it as a random variable with arbitrary absolutely continuous mixing distribution $Q_{X,Z}$ which may depend on $(X, Z)$. Define $Q$ to be the mapping from $\text{support}(X, Z)$ to the space of absolutely continuous distributions $Q(x, z) = Q_{x,z}$. The mixture model $\Pr(Y|X, Z; \beta, Q)$ for the joint distribution of $(Y|X, Z)$, marginally over $q$, is now semiparametric in that $\Pr(Y|X, q; \beta)$ is a regular parametric model while the mapping $Q$ is an infinite-dimensional parameter.

Consider a non-parametric model for the missingness process $R$ given $(Y, X, Z)$. This model is isomorphic to the set of all mappings $\lambda$ from $\text{support}(t, y_{t-1}, X, Z)$ into the unit interval such that

$$\lambda(t, y_{t-1}, x, z) = \Pr(R_s = 1|R_{s-1} = 1, Y_{t-1} = y_{t-1}, X = x, Z = z). \quad (17)$$

In an non-parametric model for $(R|Y, X, Z)$, where at least one component of $(X, Z)$ is continuous, the mapping $\lambda$ is an infinite-dimensional nuisance parameter. Assume that parameters $\lambda$ and $(\beta, Q)$ are variation independent.

For a given subject, the observed data are $(Y_{\text{obs}}, T, X, Z)$. The likelihood arising from these data, conditional on $(X, Z)$, can be written

$$\Pr(Y_{\text{obs}}, T|X, Z) = \prod_{t=1}^{T} \{\Pr(Y_t|T = t, X, Z)\Pr(T = t|X, Z)\}^{I(T=t)}$$

$$= \prod_{t=1}^{T} \{\Pr(Y_t|X, Z)\Pr(T = t|Y_t, X, Z)\}^{I(T=t)}$$

$$= \prod_{t=1}^{T} \Pr(Y_t|X, Z)^{I(T=t)} \prod_{t=1}^{T} \Pr(T = t|Y_t, X, Z)^{I(T=t)}. \quad (18)$$

The likelihood factors into a piece $L^1(\beta, Q) = \prod_{t=1}^{T} \Pr(Y_t|X, Z; \beta, Q)^{I(T=t)}$ and another piece $L^2(\lambda) = \prod_{t=1}^{T} \Pr(T = t|Y_t, X, Z; \lambda)^{I(T=t)}$. The semiparametric model for $(Y_{\text{obs}}, T|X, Z)$ is therefore a special case of the model considered by Robins, Rotnitzky and van der Laan (2000), where interest is on a subset of the parameters, in this case $\beta$, governing the factor $L^1$ of the likelihood.
By the theory outlined in Robins, Rotnitzky and van der Laan (2000), the factorization of likelihood (18) implies that the efficient score for $\beta$, in the presence of nuisance parameters ($Q, \lambda$), is orthogonal to the tangent space $W$ for the nuisance parameter $\lambda$. This result implies that, given a candidate estimating function $g(\beta, \lambda)$ for $\beta$, the efficiency of $g$ may be improved by removing from $g$ its projection onto $W$, that is, by replacing $g$ with $g - P g$ for inferences on $\beta$. For the model considered here and the estimating function $U^c$, we will show that this projection does in fact improve efficiency.

We now require a representation of the nuisance tangent space $W$ for $\lambda$ that will permit operationalization of the projection $P$. Space $W$ is the closure of the subset of $L^2$ containing scores from all parametric sub-models for the non-parametric model for $R$ indexed by $\lambda$ (Begun, Hall, Huang and Wellner, 1983). A score $S^\gamma \in W$ if (i) there exists a set of mappings

$$\lambda(t, y_{t-1}, x, z) = \lambda(t, y_{t-1}, x, z; \gamma)$$

that is a subset of (17) indexed by a finite-dimensional parameter $\gamma$, and (ii) replacing $\lambda$ with $\lambda(t; \gamma)$ in $L^2$, $S^\gamma = \partial \log L^2 \{\lambda(\cdot; \gamma)\} / \partial \gamma$ is the $\gamma$-score for the regular parametric sub-model obtained by setting $\lambda(t; \gamma) = \lambda(t; \gamma)$.

Letting $\lambda_t \equiv \lambda_t(Y_t, X, Z)$ in what follows, we can write

$$L^2(\lambda) = \prod_{t=1}^T \Pr(T = t|Y_t, X, Z; \lambda)^I(T = t) = \prod_{t=1}^T \left( \prod_{s=1}^t \lambda_s \right)^{I(T = t)} (1 - \lambda_{t+1})^{I(T = t)}$$

$$= \prod_{s=1}^J \left( \prod_{t=s}^J \lambda_s^{I(T = t)} \right) (1 - \lambda_{s+1})^{I(T = s)} = \prod_{s=1}^J \lambda_s^{I(T \geq s)} (1 - \lambda_{s+1})^{I(T = s)}$$

$$= \prod_{s=1}^J \lambda_s^{R_s(1 - \lambda_{s+1})R_{s-1}(1 - R_{s+1})} = \prod_{s=1}^J \lambda_s^{R_{s-1}R_s} \prod_{s=1}^J (1 - \lambda_{s+1})^{R_s(1 - R_{s+1})}$$

$$= \prod_{s=1}^J \lambda_s^{R_{s-1}R_s} \prod_{s=2}^{J+1} (1 - \lambda_s)^{R_{s-1}(1 - R_s)} = \prod_{s=1}^J \lambda_s^{R_{s-1}R_s} (1 - \lambda_s)^{R_{s-1}(1 - R_s)}$$

The last equality is due to the facts that $R_0 = R_1 = \lambda_1 = 1$ and $R_{J+1} = \lambda_{J+1} = 0$. 28
Finally,

\[ L^2(\lambda) = \prod_{s=1}^{J} \left\{ \lambda_s^{R_s}(1 - \lambda_s)^{(1 - R_s)} \right\}^{R_{s-1}}. \]

From this expression, it is clear that the likelihood factor \( L^2(\lambda) \) is a product likelihood arising from random variables \( R_s \) each given \((R_{s-1}, Y_{s-1}, X, Z)\). Therefore, the \( \gamma \)-scores \( S^\gamma \) for any parametric sub-model likelihood \( L^2(\lambda(\cdot; \gamma)) \) will be some linear combination of unbiased estimating functions of \((R_s, R_{s-1}, Y_{s-1}, X, Z)\) that are unbiased given \((R_{s-1}, Y_{s-1}, X, Z)\), for \( s = 1, \ldots, J \). That is, the tangent space \( \mathcal{W} \) is equal to the space spanned by \( \bigcup_{s=1}^{J} \mathcal{W}_s \).

Now, from the foregoing theory, the following facts are easily shown. First, for \( s = 1, \ldots, J \), any element of \( \mathcal{W}_s \) can be written in the form \( R_{s-1}V^s \) where \( V^s = V^s(R_s, Y_{s-1}, X, Z) \). Second, all elements of \( \mathcal{W}_s \) are orthogonal to all elements of \( \mathcal{W}_{s'} \), \( 1 \leq s < s' \leq J \). Therefore, the projection \( \mathcal{P} \) of an estimating function \( g \) is \( \mathcal{P}g = \sum_s \mathcal{P}_sg \). Finally, for \( s = 1, \ldots, J \) the projection \( \mathcal{P}_s \) of \( g \) onto \( \mathcal{W}_s \) is

\[ \mathcal{P}_sg = R_{s-1}\{E(g|R_s, R_{s-1} = 1, Y_{s-1}, X, Z) - E(g|R_{s-1} = 1, Y_{s-1}, X, Z)\}. \tag{19} \]

We now derive the projection of \( I(T = t)U^{(t)} \) onto \( \mathcal{W}_s \). First, note that \( I(T = t) = (1 - R_{t+1})R_t \). Now, consider the case where \( s \leq t + 1 \). Define

\[ V^{(t,s)}_{\text{proj}} = E\{(1 - R_{t+1})R_tU^{(t)}|R_s, R_{s-1} = 1\}, \]

and, noting that

\[ E\{(1 - R_{t+1})R_tU^{(t)}|R_{s-1} = 1\} = E[E\{(1 - R_{t+1})R_tU^{(t)}|R_s, R_{s-1} = 1\} | R_{s-1} = 1], \]

define

\[ \epsilon^{(t,s)}_{\text{proj}} = E(V^{(t,s)}_{\text{proj}}|R_{s-1} = 1). \]

Conditioning on \((Y_{s-1}, X, Z)\) is implicit in all of these expressions. So, using (19),

\[ \mathcal{P}_sI(T = t)U^{(t)} = R_{s-1}(V^{(t,s)}_{\text{proj}} - \epsilon^{(t,s)}_{\text{proj}}), \quad \text{for } s \leq t + 1. \]

Note in particular for the special case where \( s = t + 1 \), that

\[ V^{(t,t+1)}_{\text{proj}} = (1 - R_{t+1})U^{(t)} \quad \text{and} \quad \epsilon^{(t,t+1)}_{\text{proj}} = \{1 - \lambda_{t+1}(Y_{t})\}U^{(t)}. \]

29
Therefore,

\[ P_{t+1}I(T = t)U^{(t)} = R_t\{\lambda_{t+1}(Y_t) - R_{t+1}\}U^{(t)}. \]

Note also that, for \( s = t + 1 = J + 1 \), the projection is 0 because \( \lambda_{t+1}(Y_t) = R_{t+1} = 0 \).

Finally for the case where \( s > t + 1 \), \( R_{s-1} = 1 \) implies that \( R_{t+1} = R_t = 1 \). So \( I(T = t) = 0 \) regardless of the value of \( R_s \), and \( P_sI(T = t)U^{(t)} = R_{s-1}(0 - 0) = 0 \).

Alternatively, if \( R_{s-1} = 0 \), the projection is 0 by (19).

**Appendix E. Efficiency of \( U^{\text{proj}} \)**

**Proof that \( U^{\text{proj}} \) is doubly-robust.** To facilitate the argument, we expand \( U^c \) into a series of orthogonal estimating functions indexed by \( t \), each one lying in the \( L^2 \) subspace of functions of \((R_t, Y_t)\) which are unbiased conditional on \((R_{t-1}, Y_{t-1})\).

Throughout, conditioning on \((X, Z)\) is implicit, expectations are taken over \((R, Y)\), and sums are from \( t = 1 \) to \( J \). Write \( U^c = \sum_{t=1}^J g_t \), where

\[ g_t = E(U^c|R_t, Y_t) - E(U^c|R_{t-1}, Y_{t-1}). \]

Note that, for any \( g \), the projection \( P_s g \) can be written

\[ P_s g = E(g|R_s, Y_{s-1}) - E(g|R_{s-1}, Y_{s-1}). \]

It is easily shown that \( P_s g_t = 0 \) for \( s \neq t \) so that \( U^{\text{proj}} = U^c - P U^c = \sum_t (g_t - P_t g_t) \).

Now, it can be shown that

\[ P_t g_t = E(U^c|R_t, Y_{t-1}) - E(U^c|R_{t-1}, Y_{t-1}) \]

and

\[ g_t - P_t g_t = E(U^c|R_t, Y_t) - E(U^c|R_t, Y_{t-1}). \]

First, consider the case where \( \gamma^\dagger = \gamma \) and \( \alpha^\dagger \neq \alpha \). Write \( U^{\text{proj}} = U^c - \sum_t P_t g_t \), and note that unbiasedness of \( U^c \) only depends on parameters \((\beta, \gamma)\). Now, write

\[ P_t g_t = h_1(\beta, \gamma, \alpha^\dagger) - E_{R_t}\{h_1(\beta, \gamma, \alpha^\dagger)|R_{t-1}, Y_{t-1}, \gamma\}, \quad (20) \]

30
where

\[ h_1(\beta, \gamma, \alpha^\dagger) = E\{U^c(\beta, \gamma)|R_t, Y_{t-1}; \beta, \gamma, \alpha^\dagger\}. \]

The expectation in (20) is only over \( R_t \) and only depends on \( \gamma \). Therefore \( E(P_tg_t) = E(\{E_{R_t}(P_tg_t|R_{t-1}, Y_{t-1}; \gamma)\}) = E\{0\} = 0 \), completing the proof for \( \gamma^\dagger = \gamma \).

For the case where \( \alpha^\dagger = \alpha \) but \( \gamma^\dagger \neq \gamma \), write \( U^{proj} = \sum_t(g_t - P_tg_t) \). Note that

\[ g_t - P_tg_t = h_2(\beta, \gamma^\dagger, \alpha) - E_{Y_t}\{h_2(\beta, \gamma^\dagger, \alpha)|R_t, Y_{t-1}; \beta, \alpha\}, \quad (21) \]

where

\[ h_2(\beta, \gamma^\dagger, \alpha) = E\{U^c(\beta, \gamma^\dagger)|R_t, Y_t; \beta, \gamma^\dagger, \alpha\}. \]

The expectation in (21) is only over \( Y_t \) and only depends on \( (\beta, \alpha) \). Therefore \( E(g_t - P_tg_t) = E\{E_{Y_t}(g_t - P_tg_t|R_t, Y_{t-1}; \alpha)\} = E\{0\} = 0 \), completing the proof for \( \alpha^\dagger = \alpha \).

**Proof of Theorem 2.** We first demonstrate that \( U^{proj} \) is more efficient than \( U^c \). Using the expansion of \( U^c \) in the foregoing double-robustness proof, \( \text{var}(U^c) = \sum_t \text{var}(g_t) \), and \( \text{var}(U^{proj}) = \sum_t \text{var}(g_t - P_tg_t) \). We show in the following paragraph that \( \text{var}(g_t) \geq \text{var}(g_t - P_tg_t) \) in the sense that \( \text{var}(g_t) - \text{var}(g_t - P_tg_t) \) is positive semi-definite. Therefore, \( \text{var}(U^c) \geq \text{var}(U^c - P^cU^c) \). The \( \beta \)-information in \( U^{proj} \) is

\[ I_\beta^{proj} = E(-\partial U^c/\partial \beta)^T \text{var}(U^c - P^cU^c)^{-1} E(-\partial U^c/\partial \beta), \]

while that in \( U^c \) is \( I_\beta^c = E(-\partial U^c/\partial \beta)^T \text{var}(U^c)^{-1} E(-\partial U^c/\partial \beta) \). Since \( \text{var}(U^c) - \text{var}(U^c - P^cU^c) \) is positive semi-definite, the proof is complete for \( U^c \).

For showing that \( U^{proj} \) is more efficient than \( U^c \), it is left to show that \( \text{var}(g_t) \geq \text{var}(g_t - P_tg_t) \). This is accomplished by conditioning on \( (R_{t-1}, Y_{t-1}) \). Assume for ease of exposition that \( \beta \) is scalar. First, it is easily shown that \( \text{var}(P_tg_t|R_{t-1}, Y_{t-1}) = \text{cov}(g_tP_tg_t|R_{t-1}, Y_{t-1}) = \text{var}\{E(g_t|R_t, Y_{t-1})|R_{t-1}, Y_{t-1}\} \). Therefore, \( \text{var}(g_t - P_tg_t|R_{t-1}, Y_{t-1}) = \text{var}(g_t|R_{t-1}, Y_{t-1}) + \text{var}(P_tg_t|R_{t-1}, Y_{t-1}) - 2\text{cov}(g_tP_tg_t|R_{t-1}, Y_{t-1}) = \text{var}(g_t|R_{t-1}, Y_{t-1}) - \text{var}\{E(g_t|R_t, Y_{t-1})|R_{t-1}, Y_{t-1}\} \) so that \( \text{var}(g_t - P_tg_t|R_{t-1}, Y_{t-1}) \leq \text{var}(g_t|R_{t-1}, Y_{t-1}) \).
Now, since \( g_t \) and \( P_t g_t \) are unbiased given \((R_{t-1}, Y_{t-1})\), \( \text{var}(g_t) = E\{\text{var}(g_t|R_{t-1}, Y_{t-1})\} \), and similarly for \( g_t - P_t g_t \). Therefore, \( \text{var}(g_t) \geq \text{var}(g_t - P_t g_t) \).

Now to show optimality of \( U^{\text{proj}} \) in class \( U \) defined in (9), consider arbitrary \( U = U^c - U^a \in U \). By an argument analogous to the foregoing, replacing \( U^c \) with \( U \), \( U - P U \) is more efficient than \( U \), where the \( U \) information \( \mathcal{I}_\beta = \mathcal{I}_\beta^c E(UU^T)^{-1} \mathcal{I}_\beta^c \).

Finally, because all \( U^a \) defined in (7) are elements of \( \mathcal{W} \), \( U - P U = U^c - P U^c = U^{\text{proj}} \), completing the proof of optimality.

References


Table 1. Simulation results based on 1000 replicates. Upper entries are for $\beta_t$ and lower entries are for $\beta_x$. True values are $\beta_t = \beta_x = 0.405$.

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

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I(Year = 2)</td>
<td>-0.030</td>
<td>(0.07)</td>
<td>-0.20</td>
<td>(0.07)</td>
<td>-0.16</td>
<td>(0.07)</td>
</tr>
<tr>
<td>I(Year = 5)</td>
<td>0.92</td>
<td>(0.11)</td>
<td>0.66</td>
<td>(0.10)</td>
<td>0.63</td>
<td>(0.10)</td>
</tr>
<tr>
<td>I(Year = 7)</td>
<td>1.44</td>
<td>(0.14)</td>
<td>1.04</td>
<td>(0.14)</td>
<td>0.95</td>
<td>(0.14)</td>
</tr>
<tr>
<td>Age × year</td>
<td>0.11</td>
<td>(0.021)</td>
<td>0.095</td>
<td>(0.020)</td>
<td>0.096</td>
<td>(0.020)</td>
</tr>
<tr>
<td>Sex × year</td>
<td>0.047</td>
<td>(0.022)</td>
<td>0.058</td>
<td>(0.022)</td>
<td>0.064</td>
<td>(0.022)</td>
</tr>
<tr>
<td>Education × year</td>
<td>0.059</td>
<td>(0.012)</td>
<td>0.045</td>
<td>(0.012)</td>
<td>0.045</td>
<td>(0.012)</td>
</tr>
<tr>
<td>Memory × year</td>
<td>-0.033</td>
<td>(0.012)</td>
<td>-0.028</td>
<td>(0.012)</td>
<td>-0.024</td>
<td>(0.012)</td>
</tr>
</tbody>
</table>

Est., \( \hat{\beta} \) 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.