This article was downloaded by: [Virginia Tech Libraries]
On: 17 July 2013, At: 04:07
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data

James M. Robins [a] , Andrea Rotnitzky [a] & Lue Ping Zhao [b]
[a] Department of Biostatistics, Harvard University, Boston, MA, 02115
[b] Department of Epidemiology, Fred Hutchinson Cancer Center, Seattle, WA, 98104
Published online: 27 Feb 2012.

To cite this article: James M. Robins , Andrea Rotnitzky & Lue Ping Zhao (1995) Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data, Journal of the American Statistical Association, 90:429, 106-121

To link to this article: http://dx.doi.org/10.1080/01621459.1995.10476493

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://amstat.tandfonline.com/page/terms-and-conditions

Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data

James M. ROBINS, Andrea ROTNITZKY, and Lue Ping ZHAO*

We propose a class of inverse probability of censoring weighted estimators for the parameters of models for the dependence of the mean of a vector of correlated response variables on a vector of explanatory variables in the presence of missing response data. The proposed estimators do not require full specification of the likelihood. They can be viewed as an extension of generalized estimating equations estimators that allow for the data to be missing at random but not missing completely at random. These estimators can be used to correct for dependent censoring and nonrandom noncompliance in randomized clinical trials studying the effect of a treatment on the evolution over time of the mean of a response variable. The likelihood-based parametric G-computation algorithm estimator may also be used to attempt to correct for dependent censoring and nonrandom noncompliance. But because of possible model misspecification, the parametric G-computation algorithm estimator, in contrast with the proposed weighted estimators, may be inconsistent for the difference in treatment-arm–specific means, even when compliance is completely at random and censoring is independent. We illustrate our methods with the analysis of the effect of zidovudine (AZT) treatment on the evolution of mean CD4 count with data from an AIDS clinical trial.

KEY WORDS: AIDS; Counterfactual causality; Dependent censoring; G-computation algorithm; Generalized estimating equations; Intent to treat; Longitudinal data; Noncompliance; Randomized trial; Surrogate markers; Time-dependent covariates.

Downloaded by [Virginia Tech Libraries] at 04:07 17 July 2013

1. INTRODUCTION

In both randomized and nonrandomized follow-up studies, it is often of interest to estimate the evolution over time of the mean of an outcome variable of interest for the ith subject Y_{it}, $t = 1, \ldots, T$, as a function of explanatory variables X_i. In randomized studies X_i typically includes subject i's treatment arm indicator and may also include his or her baseline characteristics such as age, sex, and pretreatment clinical status. The goal of this article is to provide methods for estimating the parameters β_0 of models for the regression of Y_{it} on X_i when some Y_{it} are regarded as censored (i.e. missing), either because subjects do not comply with their assigned protocols or because they drop out of the study prior to the end of follow-up.

As discussed further in Section 2, in practice the probability that a subject is missing at the tth occasion may depend on X_i, on past values both of the outcome variable Y_{ij} and of a vector of time-dependent covariates $V_{ij}, j = (0, \ldots, t-1)$. In such cases it is well known that fully parametric likelihood methods can provide valid inferences concerning the parameters β_0 of the regression of $Y_{it}, t = 1, \ldots, T$, on X_i, if a model for the joint distribution of Y_{it}, X_i and $V_{it}, t = 0, \ldots, T$, is correctly specified, and the probability of nonresponse at t does not depend on (Y_{ij}, V_{ij}) for $j \geq t$ (Rubin 1976). But with incomplete data, likelihood methods can be sensitive to model misspecification, because they implicitly impute the missing data from their conditional distribution given the observed data (Dempster, Laird, and Rubin 1977). In addition, even with complete data, if the focus is on models for the marginal distribution of the response, then fully parametric models for certain non-Gaussian data that preserve the marginal expectation of Y_{it} given X_i can often be cumbersome and computationally difficult when X_i is multivariate with continuous components (Prentice 1988).

Liang and Zeger (1986) proposed a class of generalized estimating equations (GEE) whose solutions are consistent for β_0 whenever provided only that the model for the marginal means of the outcomes at each occasion is correctly specified. Their approach is an extension of quasi-likelihood methods (McCullagh and Nelder 1989) to the multivariate regression setting and results in iteratively reweighted least squares estimators of β_0. Similar estimators were also considered by Gourieroux, Monfort, and Trognon (1984). As Liang and Zeger pointed out, inferences with the GEE are valid only under the stronger assumption that the data are missing completely at random; that is, given X_i, the nonresponse process is independent of both observed and unobserved Y_{it}'s and V_{it}'s.

In this article we propose a class of weighted estimating equations that lead to consistent and asymptotically normal estimators of β_0 provided that the probability of nonresponse at time t, given both the entire past $\bar{W}_{it} = \{X_i, V_{i0}, Y_{i0}, \ldots, V_{i(t-1)}, Y_{i(t-1)}\}$ and the future Y's, that is, $\{Y_{it}, \ldots, Y_{iT}\}$, depends only on $\bar{W}_{it}$ and that one can connectly specify a model for the probability of nonresponse given $\bar{W}_{it}$ (Rosenbaum 1987). In contrast to the likelihood approach, our methods are computationally simple and do not require the specification of the joint distribution of the data.

In Section 2 we provide the statistical model underlying our methods and describe how it can be applied to model randomized trials of the effect of treatment on the evolution of the mean of a response variable $Y_{it}, t = 1, \ldots, T$, in the presence of dependent censoring and nonrandom noncompliance. In Section 3 we present the proposed class of estimators and discuss efficiency issues. In Section 4 we illustrate our methods with the analysis of data from an AIDS clinical trial. In Section 5 we compare our estimators to the GEE estimator and the sweep (i.e., the normal theory maximum

* James M. Robins is Professor of Epidemiology and Biostatistics and Andrea Rotnitzky is Assistant Professor, Department of Biostatistics, Harvard University, Boston, MA 02115. Lue Ping Zhao is Associate Professor, Department of Epidemiology, Fred Hutchinson Cancer Center, Seattle, WA 98104. The authors are indebted to Margaret Fischl for making the data from ACTG trial 002 available. Support for this research was provided in part by Grants 2 P30 ES00002, R01-AI32475, R01-ES03405, GM-48704, K04-ES00180, and GM-29745 from the National Institutes of Health. Andrea Rotnitzky was additionally supported in part by a Mellon Foundation Faculty Development Award.

© 1995 American Statistical Association

Downloaded by [Virginia Tech Libraries] at 04:07 17 July 2013

likelihood) estimator (Little and Rubin 1987) in a simulation study. In addition, we contrast our estimator with the likelihood-based parametric G-computation algorithm estimator (Robins 1986) and with a nonlinear generalization of the sweep estimator. In Sections 1–5 we restrict attention to monotone missing data patterns; that is, we assume that once a subject leaves the study, return is not possible. In Section 6 we generalize our results to allow for arbitrary patterns of missing data. Finally, in Section 7 we provide a summary and discuss some further considerations. In particular, we consider the possibility that our fundamental assumption that the probability of nonresponse at t is independent of future Y's given the past $\bar{W}_{it}$ is false and suggest conducting a sensitivity analysis in which one estimates the regression of Y_{it} on X_i under varying modeling assumptions about the true but nonidentifiable dependence of the nonresponse probabilities on the future Y's. The weighted estimators proposed here are analogous to estimators first proposed by Robins and Rotnitzky (1992), except that Robins and Rotnitzky allowed dropout (censoring) to occur in continuous time. Shortly thereafter, while this article was under revision for publication, Heyting, Tolboom, and Essers (1992) also proposed the use of weighted estimators for monotone missing data patterns.

2. THE MODEL

We consider a follow-up study conducted over a fixed interval from time 1 to T. Let $Y_i = (Y_{i0}, Y_{i1}, \ldots, Y_{iT})^T$ be the vector of outcome variables corresponding to the ith subject, $i = 1, \ldots, n$, measured at prespecified visit times $(0, 1, 2, \ldots, T)$, where visit 0 occurs at a time just prior to start of follow-up. Here and throughout, T used as a superscript denotes matrix transposition. Let $X_i = (X_{i0}^T, \ldots, X_{iT}^T)^T$ where X_{it} is a vector of explanatory variables associated with Y_{it} and includes the constant 1 as a component. The X_{it}, (a) may consist solely of a vector of baseline explanatory variables X_i^* in which case $X_{it} = X_i^*$, $t = 0, \ldots, T$, (b) may be a deterministic function of time and the baseline variables, e.g., $X_{it} = X_i^* t$, or (c) may be an ancillary external time varying covariate process (Kalbfleisch and Prentice 1980), such as the level of air pollution at time t in subject i's city of residence or the exposure of subject i to environmental tobacco smoke at time t as recorded by a nicotine dosimeter. The case in which the X_{it} are not ancillary is considered briefly in Section 7 and in Robins (1994a). We shall suppose that at each occasion (visit) t, $t = 1, \ldots, T$, the marginal distribution of Y_{it} given X_i follows the regression model

$$E(Y_{it}|X_i) = g_t(X_i, \beta_0) \tag{1}$$

for $i = (1, \ldots, n)$, where β_0 is a $p \times 1$ vector of unknown parameters and $g_t(\cdot, \cdot)$ are fixed functions. Our goal is to estimate β_0 when the full vector Y_i is not always observed because some subjects miss one or more visits. We assume that the study is designed so that, in addition to the Y_{it} and X_i, measurements are to be made on a vector of time-dependent covariates V_{it}, $t = (0, \ldots, T)$ and we set $W_{it} = (V_{it}^T, Y_{it})^T$, $t = (1, \ldots, T)$. Let $W_{i0} = (X_i, V_{i0}, Y_{i0})$ be comprised of X_i and baseline values Y_{i0} and V_{i0}; this notation for W_{i0} will be redundant, although accurate, if Y_{i0} and V_{i0} are already included in X_i, as would be the case if the investigator chose to adjust for baseline levels of the outcome and the time-dependent covariates in the model (1).

Define $R_{it} = 1$ if subject i is observed at time t (i.e., if Y_{it} and V_{it} are observed) and $R_{it} = 0$ otherwise. We assume that at each t, Y_{it} and V_{it} are either both observed or both missing. Throughout we assume $R_{i0} = 1$ for all subjects i. Until Section 6, we shall assume that the missing data patterns are monotone; that is, once a subject leaves the study, return is not possible, or, equivalently, $R_{it} = 0$ implies $R_{i(t+1)} = 0$. Until Section 7, we shall assume that X_i is completely observed (i.e., known); that is, X_{it} is observed even if $R_{it} = 0$. This assumption would be true for the air pollution example if subjects lost to follow-up remain in their city of residence. It will not be true for the environmental tobacco smoke example, because the environmental tobacco smoke exposure of subjects lost to follow-up will be unknown. We shall assume that the missing data process satisfies

$$P(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{it}, Y_i) = P(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{it}), \tag{2a}$$

where, for any vector z_{it}, $t = 1, \ldots, T$, $\bar{z}_{it} = (z_{i0}^T, z_{i1}^T, \ldots, z_{i(t-1)}^T)^T$ and $\bar{z}_{i0} \equiv 1$. Overbars are used to indicate variables that depend on the past data recorded up to but not including the current occasion. Throughout we shall assume

$$P(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{it}) > \sigma > 0, \quad t = 1, \ldots, T, \tag{2b}$$

so that for each subject i, the probability of remaining in the study is bounded away from zero. Eq. (2b) is needed to guarantee the existence of $n^{1/2}$-consistent estimators of β_0. Until the final section we shall assume that the vectors $(\bar{W}_{i(T+1)}^T, R_{i1}, \ldots, R_{iT})$, $i = 1, \ldots, n$ are independent and identically distributed.

Under assumption (2a), among subjects observed at time $t-1$ (i.e., those with $R_{i(t-1)} = 1$) nonresponse at time t is unrelated to the current and future outcomes $Y_{it}, \ldots, Y_{iT}$, conditional on the observed past $\bar{W}_{it}$. Assumption (2a) will be satisfied, in particular, when the data are missing at random in the sense of Rubin (1976) because, with monotone missing patterns, missing at random is equivalent to

$$P(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{i(T+1)}) = P(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{it}) \tag{3}$$

(Robins, Rotnitzky, and Zhao 1994). It is important to note that we do not assume

$$E(Y_{it}|X_i, \bar{W}_{it}) = E(Y_{it}|X_i), \tag{4}$$

it is the conditional mean of Y_i given X_i alone that is of substantive interest.

Our model can be used to describe randomized trials in the presence of dependent censoring and nonrandom noncompliance. To focus attention, consider a hypothetical 12-month two-arm randomized controlled trial of subjects with rheumatoid arthritis designed to compare the effect of a new nonsteroidal anti-inflammatory agent, say drug A, to stan-

dard aspirin therapy on the evolution of the mean of a clinical measure of disease activity Y_{it} over the entire study period. Subjects are asked to return to the clinic each month to have their clinical disease activity assessed and to have various laboratory tests done. We shall suppose that during the course of the trial, a new nonsteroidal drug B becomes available to the public and that within each treatment arm, subjects with clinical or laboratory evidence of persistent disease activity during the follow-up period are more likely than other subjects to initiate therapy with drug B and/or to fail to return to clinic for further evaluation so that noncompliance and/or dropout are nonrandom. Moreover, as it is often the case in such studies, we shall suppose that the proportion of noncompliance is greater in the standard therapy arm than in the new treatment arm. To obtain some useful information regarding the benefits of treatment A compared to aspirin therapy, it is agreed to regard a subject as *censored* at time t if therapy with B has been initiated, or if a clinic visit, and thus a urine test for drug B, has been missed at or before t. We shall further suppose that, based on substantive considerations, the investigators conducting the trial are willing to assume for the purposes of analysis that, among subjects in a given treatment arm with identical clinical and laboratory results through the tth $-$ 1 visit, the decision to initiate therapy with drug B and the decision not to return to clinic in week t are unrelated to present and future measures of disease activity; that is, to the values of Y_{ij} that would have been obtained at the jth visit, $j = t, \ldots, 12$, in the absence of therapy B or dropout. Under these assumptions, if we let (a) V_{it} record the clinical and laboratory results on the ith subject at the tth clinic visit, (b) $W_{it} = (V_{it}^T, Y_{it})^T$ and X_i includes a treatment arm indicator and possibly other baseline characteristics, and (c) $R_{it} = 1$ if a patient is uncensored at t and $R_{it} = 0$ if censored, then Equation (2a) will be true. That is, the conditional probability of becoming censored at t will be independent of Y_i given $\bar{W}_{it}$ but will be dependent on Y_i given X_i alone or even given $(X_i, \bar{Y}_{it})$. Furthermore, the parameters β_0 of the regression of Y_{it} on X_i will remain the parameters of interest even if (4) is false, because a component of β_0 will represent the treatment effect that would have been seen in the absence of both dropout and noncompliance. Note that for a subject i who becomes censored at t (i.e., $R_{it} = 0$, $R_{i(t-1)} = 1$), Y_{ij}, $j \geq t$, is a counterfactual variable (Holland 1986), because it represents the clinical disease activity that would have been observed at time j if, contrary to fact, the subject had remained uncensored through time j.

Given (2b), Equation (2a) is the fundamental assumption that will allow us to identify $E[Y_{it} \mid X_i]$ from the observable random variables in the presence of nonrandom noncompliance and dependent censoring (given only X_i). Specifically, Robins (1987a, thm. AD.1) showed that (2a) implies the *G-computation algorithm formula,*

$$E[Y_{it} \mid X_i] = \int\int \cdots \int E[Y_{it} \mid \bar{w}_t, R_{it} = 1, X_i] \times \prod_{j=0}^{j=t-1} dF[w_j \mid \bar{w}_j, R_{ij} = 1, X_i] \quad (5a)$$

and the right side of the G-computation algorithm formula depends only on the joint distribution of the observable random variables. Here w_j and $\bar{w}_j$ are realizations of W_{ij} and $\bar{W}_{ij}$. According to (5a), $E[Y_{it} \mid X_i]$ is a weighted average of the $E[Y_{it} \mid \bar{w}_t, R_{it} = 1, X_i]$ with $\bar{w}_t$-specific weights $\prod_{j=0}^{j=t-1} f(w_j \mid \bar{w}_j, R_{ij} = 1, X_i)$. If at each t the decision whether to initiate therapy with drug B and/or to drop out of the study was decided at random by the flip of a coin, then (3) and thus (2a) would be true even if the probability of the coin landing heads depended on past $\bar{W}_t$ history. Because in practice noncompliance and dropout are not decided by the flip of a coin, it is a primary goal of the investigators conducting the study to record in V_t data on a sufficient number of covariates to ensure that (2a) holds, at least to a good approximation. But irrespective of the number of covariates recorded in V_t as "an observational study assumption," Equation (2a) cannot be guaranteed to hold and cannot be subjected to an empirical test.

It is of interest to note that (3) implies, but (2a) does not imply, that $f[W_{it} \mid \bar{W}_{it}, R_{it} = 1] = f[W_{it} \mid \bar{W}_{it}]$ (Robins 1987a). Nonetheless, (2a) is sufficient to imply the G-computation algorithm formula (5a) that identifies $E[Y_{it} \mid X_i]$. Plausible substantive scenarios in which (2a) is true but (3) is false can be constructed along the lines described in Robins (1987a, Sec. AD.3) and Robins (1993a, App. 2). In fact, Robins (1989, sec. 13; 1994a, thm. A.3) showed that (5a) is implied by the even weaker condition

$$E[Y_{it} \mid \bar{W}_{ij}, R_{ij} = 1] = E[Y_{it} \mid \bar{W}_{ij}, R_{i(j-1)} = 1], \qquad j \leq t. \quad (5b)$$

Note that (2a) implies (5b) but (5b) does not imply (2a).

Remark. To see that theorem A.3 of Robins (1994a) proves that (5b) implies (5a), we regard the R_{ij} as a sequence of time-dependent "treatments" indexed by j and consider estimation of the mean of Y_{it} had, contrary to fact, all subjects followed the treatment regime G specified by $R_{ij} = 1$ for $j = 0, \ldots, t$; that is, the regime in which no subject was censored. Robins (1986, p. 1491; 1987a, sec. AD.5) provided additional discussion of the usefulness of regarding censoring or missing data indicators as time-dependent treatments.

Researchers studying causal models based on directed acyclic graphs (Spirtes, Glymour, and Scheines, 1993; Pearl, 1994) have recently rediscovered the G-computation algorithm formula as an element of the so-called "manipulation theorem." Robins (1995) showed that the manipulation theorem itself is, in essence, a special case of theorem AD.1 and its corollary in Robins (1987b, pp. 929–930).

3. A CLASS OF ESTIMATORS

In this section we propose a class of estimating equations for β_0. Let $\varepsilon_{it}(\beta) = Y_{it} - g_t(X_i, \beta)$ and $\varepsilon_i(\beta) = (\varepsilon_{i1}(\beta), \ldots, \varepsilon_{iT}(\beta))^T$, and let $D_i(\beta) = d(X_i, \beta)$ be a $p \times T$ matrix of fixed functions of X_i and β. Define $H_i(\beta) = D_i(\beta)\varepsilon_i(\beta)$. When Y_i is observed for all subjects, then, under mild regularity conditions, the estimating equation

$$U_{\text{full}}(\beta) \equiv n^{-1/2} \sum_{i=1}^{n} H_i(\beta) = 0$$

Downloaded by [Virginia Tech Libraries] at 04:07 17 July 2013

has a root that is consistent for β_0 and asymptotically normal (Liang and Zeger 1986). When some of the Y_{it} are missing and nonresponse is independent of the outcomes given X_i, that is,

$$P(R_{it} = 1 \mid R_{i(t-1)} = 1, X_i, Y_i) = P(R_{it} = 1 \mid R_{i(t-1)} = 1, X_i), \quad (6)$$

then the estimating equation based on the observed data, $U_{\text{comp}}(\beta) = n^{-1/2} \sum_{i=1}^{n} D_i^*(\beta)\varepsilon_i^*(\beta) = 0$, has, under regularity conditions, a root that is consistent for β_0 and that is asymptotically normally distributed. Here for each β, $\varepsilon_i^*(\beta) = Y_i^* - g^*(X_i, \beta)$ is equal to the vector of observed residuals for subject i and $D_i^*(\beta)$ is any conformable matrix of functions of X_i and β with p rows. The class of estimating equations $U_{\text{comp}}(\beta)$ includes the GEE proposed by Liang and Zeger (1986). In the GEE, $D_i^*(\beta) = \{\partial g^*(X_i, \beta)/\partial\beta\} c^*(X_i)^{-1}$, where $c^*(X_i)$ is the conformable upper left square submatrix of an arbitrary $T \times T$ "working" covariance matrix $c(X_i)$ chosen by the investigator. When (6) is false but (2a) holds, the solution of $U_{\text{comp}}(\beta) = 0$ will typically fail to be consistent, since $\varepsilon_i^*(\beta_0)$ will no longer have mean zero conditional on response (Liang and Zeger 1986) because, as noted by Rubin (1976), subjects with $R_{it} = 1$ will represent a biased sample. Note that given (2a), (6) is implied by the testable restriction

$$P(R_{it} = 1 \mid R_{i(t-1)} = 1, X_i, \bar{W}_{it}) = P(R_{it} = 1 \mid R_{i(t-1)} = 1, X_i). \quad (7)$$

Although (6) may logically be true even if (2a) and (7) are both false, from a substantive viewpoint it might be unlikely for an investigator to believe that (6) was true unless he or she also believed that both (2a) and (7) were true. We shall suppose that the response probabilities $\bar{\lambda}_{it} = P(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{it})$ are known up to a $q \times 1$ vector of unknown parameters α_0. That is, we assume that there exists $\bar{\lambda}_{it}(\alpha)$, a known function of α and $\bar{W}_{it}$ taking values on (0, 1], such that

$$\bar{\lambda}_{it} = \bar{\lambda}_{it}(\alpha_0). \quad (8)$$

Typically, $\bar{\lambda}_{it}(\alpha)$ would be chosen to be a logistic function. That is, one would assume that given $R_{i(t-1)} = 1$, R_{it} follows a logistic model on functions of $\bar{W}_{it}$ indexed by α. Standard procedures can be used to investigate the functional form of $\bar{\lambda}_{it}(\cdot)$, because the response probability $P(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{it})$ depends only on the observed $\bar{W}_{it}$'s. Let $\hat{\alpha}$ be the partial maximum likelihood estimator (MLE) of α_0 in model (8); that is, $\hat{\alpha}$ maximizes the partial likelihood,

$$L(\alpha) = \prod_i L_i(\alpha) = \prod_i \prod_t [\bar{\lambda}_{it}(\alpha)^{R_{it}}\{1 - \bar{\lambda}_{it}(\alpha)\}^{1-R_{it}}]^{R_{i(t-1)}}. \quad (9)$$

Let $S_{\alpha i}(\alpha) = \{\partial \log L_i(\alpha)/\partial\alpha\} = \sum_{t=1}^{T} (R_{it} - \bar{\lambda}_{it}(\alpha) R_{i(t-1)})\{\partial \text{ logit } \bar{\lambda}_{it}(\alpha)/\partial\alpha\}$ be the contribution to the score for α from the ith subject and let $\Omega = [\text{var}\{S_{\alpha i}(\alpha_0)\}]^{-1}$ be the asymptotic variance of $n^{1/2}(\hat{\alpha} - \alpha_0)$. Note that $S_{\alpha i}(\alpha)$ simplifies to $\sum_{t=1}^{T} (R_{it} - \bar{\lambda}_{it}(\alpha) R_{i(t-1)}) h(\bar{W}_{it})$ if $\bar{\lambda}_{it}(\alpha)$ follows the logistic regression model logit $\bar{\lambda}_{it}(\alpha) = \alpha^T h(\bar{W}_{it})$ for some $q \times 1$ known vector function $h(\cdot)$. Define $\bar{\pi}_{it}(\alpha) = \bar{\lambda}_{i1}(\alpha) \times \cdots \times \bar{\lambda}_{it}(\alpha)$. When (3) holds, $\bar{\pi}_{it}(\alpha_0)$ is the conditional probability of observing subject i at the tth occasion given the entire vector $\bar{W}_{i(T+1)}$ (Rubin 1976). Let

$$\Delta_i(\alpha) = \begin{bmatrix} \bar{\pi}_{i1}(\alpha)^{-1} R_{i1} & 0 & \cdots & 0 \\ 0 & \bar{\pi}_{i2}(\alpha)^{-1} R_{i2} & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \bar{\pi}_{iT}(\alpha)^{-1} R_{iT} \end{bmatrix}$$

be the $T \times T$ diagonal matrix with diagonal elements $\Delta_{it}(\alpha) = \bar{\pi}_{it}(\alpha)^{-1} R_{it}$. Let $U_i(\beta, \alpha) = D_i(\beta)\Delta_i(\alpha)\varepsilon_i(\beta)$. Notice that premultiplying $\varepsilon_i(\beta)$ by $\Delta_i(\alpha)$ is tantamount to weighting each component $\varepsilon_{it}^*(\beta)$ of the vector of observed residuals $\varepsilon_i^*(\beta)$ by $\bar{\pi}_{it}(\alpha)^{-1}$. Consider the estimating equations

$$U(\beta, \hat{\alpha}) \equiv n^{-1/2} \sum_{i=1}^{n} U_i(\beta, \hat{\alpha}) = 0. \quad (10)$$

Weighted estimating equations have also been considered by Horvitz and Thompson (1952), Manski and Lerman (1977), Kalbfleisch and Lawless (1988), Rosenbaum (1987), Flanders and Greenland (1991), Zhao and Lipsitz (1992), Robins and Rotnitzky (1992), and Robins et al. (1994). Theorem 1 states that (10) has a root $\hat{\beta}$ that is consistent and asymptotically normal for estimating β_0. Consistency of $\hat{\beta}$ follows from the fact that $U_i(\beta_0, \alpha_0)$ has mean zero under (1), (2), and (8). To help understand why, note again that when the stronger assumption (3) is true, $\bar{\pi}_{it}(\alpha_0)$ is the conditional probability given $\bar{W}_{i(T+1)}$ that subject i remains uncensored through occasion t. Hence if subject i has $R_{it} = 1$ and $\bar{\pi}_{it} = .1$, then subject i would, as in $U_i(\beta_0, \alpha_0)$, need to account at time t for $1/\bar{\pi}_{it}(\alpha_0) = 10$ subjects (himself and 9 other similar subjects who were censored prior to t). Under the weaker assumptions (2a) or (5b), $\bar{\pi}_{it}(\alpha_0)$ can no longer be interpreted as the conditional probability that subject i remains uncensored through time t; nevertheless, as proved in Lemma A.1 of Appendix A, the G-computation algorithm formula (5a) implies (and is implied by) $E[Y_{it}R_{it}/\bar{\pi}_{it} \mid X_i] = E[Y_{it} \mid X_i]$ with $\bar{\pi}_{it} = \bar{\pi}_{it}(\alpha_0)$, and thus $U_i(\beta_0, \alpha_0)$ still has mean zero. For notational convenience, let $G_i(\beta) = (g_1(X_i, \beta), \ldots, g_T(X_i, \beta))^T$. Let $\Lambda_i(\alpha, \beta) = \text{Diag}\{\Delta_{it}(\alpha)^2 \varepsilon_{it}(\beta)\}$ denote the $T \times T$ diagonal matrix with diagonal elements $\Delta_{it}(\alpha)^2 \varepsilon_{it}(\beta)$.

Theorem 1. Subject to the regularity conditions stated in Appendix A, if (1), (2a), (2b), and (8) are true, then, (a) with probability approaching 1, there exists a unique solution $\hat{\beta}$ to (10); and (b) $U(\beta_0, \hat{\alpha})$ and $n^{1/2}(\hat{\beta} - \beta_0)$ are asymptotically normal with mean zero and respective asymptotic variances C and $\Gamma^{-1}C\Gamma^{-1,T}$ that can be consistently estimated by $\hat{C}$ and $\hat{\Gamma}^{-1}\hat{C}\hat{\Gamma}^{-1,T}$, where $\Gamma = E\{\partial U_i(\beta_0,$

Downloaded by [Virginia Tech Libraries] at 04:07 17 July 2013

$\alpha_0)/\partial\beta^T\} = E\{\partial H_i(\beta_0)/\partial\beta^T\}$, $C = I - B\Omega B^T$, $I = E\{U_i(\beta_0, \alpha_0)U_i(\beta_0, \alpha_0)^T\}$, $B = E\{\partial U_i(\beta_0, \alpha_0)/\partial\alpha^T\}$, $\hat{\Gamma} = n^{-1/2}\partial U(\hat{\beta}, \hat{\alpha})/\partial\beta^T = n^{-1}\Sigma D_i(\hat{\beta})\Delta_i(\hat{\alpha})\{\partial G_i(\hat{\beta})/\partial\beta^T\}$, and $\hat{C} = \hat{I} - \hat{B}\hat{\Omega}\hat{B}^T$ with $\hat{I} = n^{-1}\Sigma U_i(\hat{\beta}, \hat{\alpha})U_i(\hat{\beta}, \hat{\alpha})^T$, $\hat{B} = n^{-1/2}\partial U(\hat{\beta}, \hat{\alpha})/\partial\alpha^T = n^{-1}\Sigma D_i(\hat{\beta})\Lambda_i(\hat{\alpha}, \hat{\beta})\{\partial\bar{\pi}_i(\hat{\alpha})/\partial\alpha^T\}$, and $\hat{\Omega}$ is the estimate of Ω based on the observed information from the partial likelihood (9).

Robins and Rotnitzky (1995) show that the semiparametric model for the observables characterized by restrictions (1), (2a), (2b), and (8) is identical to that characterized by (1), (2b), (5b), and (8). Hence, Theorem 1 remains true with Equation (5b) substituted for Equation (2a).

The proofs of Theorem 1 and all subsequent theorems are given in the Appendix. The variance estimator given in Theorem 1 may not be positive definite. But we can obtain a positive definite variance estimator as follows. As is shown in the proof of Theorem 1, $U(\beta, \hat{\alpha})$ and $\hat{\beta}$ admit the asymptotic expansions

$$U(\beta, \hat{\alpha}) = n^{-1/2}\Sigma_i \text{Resid}(U_i, S_{\alpha,i}) + o_p(1),$$

$$n^{1/2}(\hat{\beta} - \beta_0) = -\Gamma^{-1}n^{-1/2}\Sigma_i \text{Resid}(U_i, S_{\alpha,i}) + o_p(1), \quad (11)$$

where $U_i = U_i(\beta_0, \alpha_0)$, $S_{\alpha,i} = S_{\alpha,i}(\alpha_0)$, $\text{Resid}(A_i, B_i) = A_i - E[A_iB_i^T]\{E(B_iB_i^T)\}^{-1}B_i$ is the residual from the population least squares regression of A_i on B_i. Hence $C = \text{var}\{\text{Resid}(U_i, S_{\alpha,i})\}$. It follows from (11) that a consistent and positive definite estimator of the asymptotic variance of $n^{1/2}(\hat{\beta} - \beta_0)$ is $\hat{\Gamma}^{-1}\hat{C}\hat{\Gamma}^{-1T}$, where

$$\hat{C} = n^{-1}\Sigma_i[\hat{\text{R}}\text{esid}\{\hat{U}_i, \hat{S}_{\alpha,i}\}]^{\otimes 2},$$

$A^{\otimes 2} = AA^T$, $\hat{U}_i = U_i(\hat{\beta}, \hat{\alpha})$, $\hat{S}_{\alpha,i} = S_{\alpha,i}(\hat{\alpha})$, and $\hat{\text{R}}\text{esid}(A_i, B_i) = A_i - (\Sigma_i A_i B_i^T)(\Sigma_i B_i B_i^T)^{-1}B_i$ is the residual for subject i from the least squares regression of A_i on B_i, $i = 1, \ldots, n$. That is, $\hat{C}$ is equal to $1/n$ times the matrix of sum of squares and cross-products of the residuals from the multivariate regression of the $p \times 1$ vectors $\hat{U}_i$ on the $q \times 1$ vectors $\hat{S}_{\alpha,i}$, $i = 1, \ldots, n$, and thus it can be computed using standard linear models software.

Given a sequence of J correctly specified nested models $\bar{\lambda}_{it}^{(j)}(\cdot), j = 1, \ldots, J$, for $\bar{\lambda}_{it}$ ordered by increasing dimension of their parameter vector $\alpha^{(j)}$, the asymptotic variance matrix of the estimator $\hat{\beta}^{(j)}$ based on the jth model will never increase and will typically decrease as the dimension of $\alpha^{(j)}$ increases. This follows from the fact that (a) the variance matrix of the residuals from a population multivariate regression decreases as the number of covariates increases and the score $S_{\alpha^{(j)},i}$ for the jth model has the same dimension as $\alpha^{(j)}$, and (b) Γ does not depend on the model $\bar{\lambda}_{it}^{(j)}(\cdot)$. Here for matrices A and B, $A \geq B$ if only if $A - B$ is nonnegative definite. Thus augmenting the selection model (8) will usually lead to an improvement of the efficiency with which we estimate β_0. We can construct an augmented selection model by, for example, adding to a linear logistic model in the components of $\bar{W}_{it}$ additional terms, such as powers of the continuous components of $\bar{W}_{it}$ and interaction terms between some of the components of $\bar{W}_{it}$. Using standard expansions for M estimators, it can be shown that if $\bar{\lambda}_{it}$, $t = 1, \ldots, T$ are known and if the true values $\bar{\lambda}_{it}$ are used instead of $\bar{\lambda}_{it}(\hat{\alpha})$ in (10), then Equation (10) has a root $\hat{\beta}^\dagger$ that satisfies $n^{1/2}(\hat{\beta}^\dagger - \beta_0) = -\Gamma^{-1}n^{-1/2}\sum_{i=1}^n U_i + o_p(1)$. But $\text{var}\{\text{Resid}(U_i, S_{\alpha,i})\} \leq \text{var}(U_i)$, because the variance of U_i is never smaller than the residual variance of the population regression of U_i on $S_{\alpha,i}$. Thus, even when the nonresponse probabilities $\bar{\lambda}_{it}$ are known, the estimator $\hat{\beta}$ that uses the estimated probabilities $\bar{\lambda}_{it}(\hat{\alpha})$ is at least as efficient as $\hat{\beta}^\dagger$, the estimator that uses the true $\bar{\lambda}_{it}$.

We have found that augmenting a correctly specified selection model usually leads to an improvement in the efficiency with which we estimate β_0. But there is a practical consideration regarding the degree of parameterization of the nonresponse model. Our proof of Theorem 1 requires that $\hat{\alpha} - \alpha_0 = O_p(n^{-1/2})$. Newey (1993) noted that it should be sufficient for $\hat{\alpha}$ to be $n^{1/4+\delta}$ consistent for $\delta > 0$ in order that $n^{1/2}(\hat{\beta} - \beta_0)$ be asymptotically normal and unbiased, which places limits on the number of free parameters in $\bar{\lambda}_{it}(\alpha)$. Further, as stated in the following lemma, there exists a lower bound such that no solution of (10) can have an asymptotic variance smaller than this bound, regardless of the degree of parameterization of the model chosen for $P(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{it})$.

Lemma 1. For fixed choice of $d(\cdot, \cdot)$, $\text{var}^A[n^{1/2}(\hat{\beta} - \beta_0)] \geq \Gamma^{-1}\text{var}(U_i - S_i)\Gamma^{-1,T}$, where $S_i = \sum_{t=1}^T (R_{it} - \bar{\lambda}_{it}(\alpha_0)R_{i(t-1)})\bar{Q}_{it}$ with equality if and only if $b\partial \text{ logit } \bar{\lambda}_{it}(\alpha_0)/\partial\alpha = \bar{Q}_{it}$, $t = 1, \ldots, T$, for some matrix b. Here $\bar{Q}_{it} = D_i(\beta_0)\bar{Q}_{it}^*$, where $\bar{Q}_{it}^*$ is the T vector with the first $t - 1$ components zero and, for $t \leq m \leq T$, the mth component $\bar{\pi}_{it}^{-1}E\{\varepsilon_{im}(\beta_0) \mid R_{i(t-1)} = 1, \bar{W}_{it}\}$.

Lemma 1 implies that for a fixed choice of $d(\cdot, \cdot)$, once a linear logistic term in $\bar{Q}_{it}$ has been included in a correctly specified model $\bar{\lambda}_{it}(\alpha)$, we no longer increase the efficiency of $\hat{\beta}$ by adding additional covariates. Because the class of estimators $\hat{\beta}$ solving (10) were proposed on an ad hoc basis, one might be concerned that these estimators fail to extract all the available information concerning β_0 contained in our assumptions (1), (2), and (8). But Robins and Rotnitzky (1995) prove that there exists an estimator $\hat{\beta}_{op}$ in our class whose asymptotic variance attains the semiparametric variance bound for our model. That is, there is no regular estimator of β_0 that is guaranteed to be asymptotically normal and unbiased under the sole restrictions that (1), (2), and (8) are true with asymptotic variance less than that of $\hat{\beta}_{op}$.

4. A WORKED EXAMPLE

We use data from the AIDS Clinical Trial Group Protocol 002 (Fischl et al. 1990) that compared high-dose versus low-dose zidovudine (AZT) treatment on the survival of AIDS patients. Measurements for this study were obtained on CD4 lymphocyte counts at baseline and at clinic visits occurring approximately at weeks 8, 16, 24, and 32. Our goal is to estimate the evolution of the mean of CD4 count differences between the two treatment arms that would have been observed in the absence of nonresponse. In this analysis we regard a subject as missing (i.e., censored), once a clinic visit is missed, and we ignore all subsequent data on that subject.

Downloaded by [Virginia Tech Libraries] at 04:07 17 July 2013

Downloaded by [Virginia Tech Libraries] at 04:07 17 July 2013

For the purpose of this analysis only, we treat censoring as independent of subsequent CD4 count given past CD4 count and total white blood cell count history. That is, we assume that Equation (2) holds with V_{it} recording total white blood cell count and Y_{it} recording CD4 count for the ith subject at the tth visit, where visit t refers to the clinic visit at week $8t$, $t = 0, 1, 2, 3, 4$. Furthermore, we regard nonresponse due to death as no different from other types of nonresponse. Our analysis is for illustrative purposes only. A more complete analysis would need to determine other potential determinants of nonresponse; further it would model nonresponse due to death separately from nonresponse due to other causes. See Robins (1993b, p. 27) for a worked example in which different forms of nonresponse are separately modeled.

Rows 1 and 2 of Table 1 record the number of uncensored subjects with CD4 count measurements at each follow-up time. At baseline there were 259 subjects in the low-dose arm and 260 subjects in the high-dose arm. Thereafter, censoring in the high-dose arm exceeded that in the low-dose arm. (One subject with extreme outliers in his CD4 measurements was excluded.) We estimated the treatment arm–specific mean of CD4 count and the mean difference at each follow-up time. In rows 3–11 our results are based on estimating the mean CD4 count $\beta_{0,x,t}$ at the tth occasion, $t = 1, \ldots, 4$, in treatment arm x by

$$\hat{\beta}_{x,t} = \sum_{i=1}^{n} I(X_i = x)\bar{\pi}_{it}(\hat{\alpha})^{-1}R_{it}Y_{it} \div \sum_{i=1}^{n} I(X_i = x)\bar{\pi}_{it}(\hat{\alpha})^{-1}R_{it}, \quad (12)$$

where X_i is a dichotomous treatment arm indicator taking the value $x = 1$ in the high-dose arm and $x = 0$ in the low-dose arm and $I(X_i = x)$ is equal to 1 if $X_i = x$ and 0 otherwise. The estimators (12) of the treatment arm–specific means are solutions to the estimating equation (10) with $g_t(X_i, \beta_0) = \sum_{x=0}^{1} I(X_i = x)\beta_{0,x,t}$ and $d(x) = \partial g(x, \beta_0)/\partial\beta$, where β_0 is the 8×1 vector with components $\beta_{0,x,t}$; $x = 0, 1$; $t = 1, 2, 3, 4$. In rows 9–11 $\hat{\alpha}$ was based on fitting separately in each treatment arm and at each occasion t a logistic model for $\bar{\lambda}_{it}$ that included an intercept and the CD4 count measurements at time $0, \ldots, t-1$ as covariates. Consequently, $\hat{\beta}_1$ and $\hat{\beta}_0$ are independent, where $\hat{\beta}_x = (\hat{\beta}_{x,1} \ldots \hat{\beta}_{x,4})^T$. In rows 6–8 we added to the logistic model covariates representing the square of CD4 count at time $t-1$ and, if $t \geq 2$, the square of CD4 count at $t-2$ and the interaction of CD4 count at $t-1$ and $t-2$. In the analysis of rows 3–5 we further added the covariates for white blood cell count at $t-1$ and the interaction between white blood cell count at $t-1$ and the CD4 count at $t-1$. Further, if $t \geq 2$, we also added white blood cell count at $t-2$. The ΔG^2 change in deviance statistic and the associated degrees of freedom for these logistic models are given in rows 21–26 of Table 1. For example, the change in deviance of the logistic model for the probability of responding at occasion 1 in the low-dose arm associated with adding the covariates corresponding to white blood cell count at baseline (i.e., occasion 0) and the interaction of white blood cell count at baseline and CD4 count at baseline to the model that included nonlinear functions of CD4 count was .1. The change in the deviance associated with adding the covariate for the square of CD4 count at baseline to the logistic model with linear CD4 count terms was 2.2. Finally, the change in the deviance associated with adding the covariates CD4 count at baseline to the logistic model that included only a constant term was 2.3.

On inspecting the results in rows 3–5 based on our most richly parameterized nonresponse model, we see that the estimate of the mean CD4 count tended to be greater in the low-dose arm than in the high-dose arm at weeks 8 and 32, although the difference was not statistically significant.

More interestingly, in comparing row 4 with row 7, we see that the estimated mean CD4 count in the high-dose arm based on the nonresponse model that included white blood cell count as a predictor is four to five cells per cubic centimeter less than the estimates based on a logistic nonresponse model that did not include white blood cell count as a predictor. This presumably reflects the fact that white blood cell count and its interaction with CD4 count are predictors both of censoring and of subsequent CD4 count. Evidence that white blood cell count and its interaction with CD4 count predict censoring can be seen in the large ΔG^2 statistic of 18.7 with 2 degrees of freedom in row 22 at visit $t = 1$. In contrast, comparing rows 6 and 9 and rows 7 and 10, we see that the estimated means of CD4 count changed little when we added nonlinear terms in CD4 count to the logistic model used in rows 9–11. The estimated standard errors in rows 3–5 are less than those in the corresponding rows 6–8. This improvement in the standard errors with increasing parameterization of the nonresponse model would be as predicted by our theoretical results if the nonresponse model in both rows 6–8 and rows 3–5 were correctly specified and if past white blood cell count and/or the interaction of past white blood cell count with past CD4 count independently predicted subsequent CD4 count conditional on past CD4 count. Because the chi-squared value of 18.7 on 2 degrees of freedom in row 22 suggests that the nonresponse model used in rows 6–8 was misspecified, improvement in the estimated standard error should not be seen as a reflection of our theory. Indeed, if the nonresponse model used in rows 6–8 is misspecified, then the standard error estimator is an inconsistent estimator of the true standard error of the possibly inconsistent $\hat{\beta}$; furthermore, neither the true standard error of $\hat{\beta}$ nor its inconsistent estimator in rows 6–8 need be larger than that of rows 3–5. As an example of this phenomenon, the estimated standard errors in rows 9–11 are not always less than those in rows 6–8, although the response model used in the latter rows is more richly parameterized.

Rows 12–14 and 15–18 report estimates of the treatment arm–specific mean CD4 counts and differences using the "sweep" estimator for monotone missing data discussed by Little and Rubin (1987, p. 112) and the "optimal" GEE estimator that uses as a working covariance matrix the empirical covariance between the Y_{it} based on the "available cases." (That is, the working covariance matrix in arm x, $x = 0, 1$, has t, t' entry $\{\Sigma R_{it}R_{it'}Y_{it}Y_{it'}/\Sigma R_{it}R_{it'}\} - \{\Sigma R_{it}Y_{it}/\Sigma R_{it}\}\{\Sigma R_{it'}Y_{it'}/\Sigma R_{it'}\}$, where the sums are over subjects in arm x). Both the sweep estimator and the optimal GEE

Table 1. Adjusted Estimates of Mean CD4 Counts

	Row	Arm	Visit number (week 8t) t = 1	t = 2	t = 3	t = 4
Persons remaining	1	low	223	187	158	120
	2	high	205	167	132	101
Nonresponse models						
WBC, CD4	3	low	147.9 (9.3)	102.7 (8.1)	70.8 (5.4)	63.7 (6.0)
	4	high	134.1 (7.3)	101.7 (9.1)	74.0 (6.5)	55.8 (5.6)
	5	diff	13.8 (11.8)	1.0 (12.2)	−3.2 (8.5)	7.9 (8.2)
Nonlinear CD4	6	low	148.0 (9.5)	103.0 (8.4)	70.9 (5.5)	63.8 (6.0)
	7	high	139.6 (8.0)	105.4 (9.3)	77.6 (6.8)	60.2 (6.2)
	8	diff	8.4 (12.4)	−2.4 (12.5)	−6.7 (8.7)	3.6 (8.6)
Linear CD4	9	low	149.8 (9.7)	104.1 (8.8)	72.3 (6.4)	65.0 (6.5)
	10	high	138.7 (7.6)	103.9 (8.8)	78.6 (6.5)	62.2 (6.5)
	11	diff	11.1 (12.3)	.2 (12.4)	−6.3 (9.1)	2.8 (9.2)
"Optimal" GEE	12	low	150.0	104.7	72.8	63.7
	13	high	138.7	103.7	78.2	62.4
	14	diff	11.3	1.0	−5.4	1.3
Sweep	15	low	151.6 (10.2)	105.6 (9.2)	73.7 (6.8)	65.8 (6.8)
	16	high	138.7 (7.7)	103.9 (8.9)	79.0 (7.0)	63.0 (7.3)
	17	diff	12.9 (12.8)	1.7 (12.8)	−5.3 (9.8)	2.8 (10.0)
Sample average	18	low	147.2 (9.6)	103.2 (9.0)	75.1 (7.5)	72.3 (8.1)
	19	high	139.0 (7.9)	107.1 (9.5)	82.1 (7.9)	68.1 (10.0)
	20	diff	8.2 (12.4)	−3.9 (13.1)	−7.0 (10.9)	4.2 (12.9)
Deviances ΔG^2 (d.f.)						
WBC/Nonlinear CD4	21	low	.1 (2)	4.6 (3)	1.1 (3)	16.2 (3)
	22	high	18.67 (2)	7.5 (3)	9.7 (3)	1.1 (3)
Nonlinear/linear CD4	23	low	2.2 (1)	3.4 (3)	.5 (3)	1.9 (3)
	24	high	1.7 (1)	4.8 (3)	10.3 (3)	2.9 (3)
Linear CD4/intercepts only	25	low	2.3 (1)	.9 (2)	3.5 (3)	7.6 (4)
	26	high	0 (1)	1.9 (2)	3.1 (3)	2.9 (4)

NOTE: Standard errors are in parentheses.

Downloaded by [Virginia Tech Libraries] at 04:07 17 July 2013

estimator ignore data on white blood cell count history. If the data are missing completely at random (i.e., the probability of nonresponse depends only on the subject i's treatment arm X_i), then the sweep estimator and the optimal GEE estimator are asymptotically equivalent. In particular, the sweep estimator of mean CD4 count is consistent when the data are missing completely at random. Little and Rubin (1987) showed that the sweep estimator is also consistent if (a) Equation (2) holds with $\bar{W}_{it}$ consisting solely of X_i and past CD4 count history Y_{it} and (b) the Y_{it}, $t = 0, \ldots, 4$, are jointly normally distributed given X_i. In Theorem B2 in Appendix B we prove that, given the foregoing condition (a), the sweep estimator is consistent and asymptotically normal if (b′) within each treatment arm, the conditional mean of Y_{it} given $\bar{Y}_{it}$ is linear in $\bar{Y}_{it}$ for each t. Furthermore, we derive a robust consistent estimator for the asymptotic variance of the sweep estimator under (a) and (b′), which are used to obtain the standard error estimates in rows 12–14. (Little and Rubin [1987] did not provide a variance estimator for the sweep estimator.) Note that condition (b′) is weaker than condition (b). But we have previously seen evidence that condition (a) is false, because past white blood cell count history predicts censoring and is correlated with future CD4 counts. Thus in our setting, the usual sweep estimator is expected to be inconsistent. Although we did not report the variance of the optimal GEE estimator, Robins and Rotnitzky (1995) show that if the data are missing completely at random, then the asymptotic variance would be equal to that of both the linear logistic estimator of rows 9–11 and the sweep estimator.

Finally, in rows 18 and 19, the treatment arm–specific estimates of the mean CD4 count at occasion t are computed as the sample averages of CD4 count measurements based on subjects that are uncensored at occasion t. If the data were missing completely at random, then these estimates would be consistent but less efficient than any of the other estimates, including the sweep and GEE estimators, because the sample averages do not exploit the fact that CD4 counts are correlated over time. The estimator based on sample averages will be inconsistent when the data are not missing completely at random.

In summary, there is some evidence in the data for dependent censoring due to the fact that a subject's white blood cell count history predicts both loss to follow-up and subsequent CD4 count.

5. A SIMULATION STUDY

In this section we report results of a small simulation study done to evaluate the performance of the weighted estimators

Table 2. Results of a Simulation Study at $t = 3$ with $\beta_{0,3} = E[Y_{i3}] = 80.0$

Row	Analysis method	Nonresponse model	Monte Carlo average of $\hat{\beta}_3$		Actual coverage rate of $\hat{\beta}_3 \pm 1.96\sqrt{\text{s.e.}\hat{\beta}_3}$ in %		Monte Carlo variance of $\hat{\beta}_3$		Monte Carlo average of $\hat{M}EP3$		Monte Carlo average of $\hat{M}IP3$	
			$\rho^2 = .81$	$\rho^2 = .36$	$\rho^2 = .81$	$\rho^2 = .36$	$\rho^2 = .81$	$\rho^2 = .36$	$\rho^2 = .81$	$\rho^2 = .36$	$\rho^2 = .81$	$\rho^2 = .36$
1	Sample average		86.4	84.2	0	4	1.2	1.2				
2	GEE (optimal)		84.2	82.5			1.2	1.2				
3	SWEEP		83.9	82.5	5	35	1.2	1.2				
4	Weighted	Linear CD4	84.0	82.5	3	44	1.5	1.3	.46	.44	.14	.15
5	Weighted	Last tercile WBC	80.0	80.0	94.5	96	1.8	1.8	.69	.69	.12	.13
6	Weighted	Linear CD4 & last tercile WBC	80.1	80.0	94	95.5	1.7	1.8	.67	.67	.10	.10
7	Weighted	Linear CD4 & last WBC	74.8	76.3	90	91	36.3	22.9	.55	.55	.02	.02
8	Weighted	Linear CD4 & last WBC & last tercile WBC	80.1	80.1	96	95.5	1.6	1.7	.67	.66	.10	.10
9	Extended Sweep	Using WBC	79.6	79.9	93.5	93.5	1.4	1.5				
10	Extended Sweep	Using $(WBC)^{1.33}$	80.0	80.0	94	95	1.3	1.4				
11	Extended GEE (optimal)	Using WBC	81.8	81.2			1.4	1.4				
12	Extended GEE (optimal)	Using $(WBC)^{1.33}$	81.4	81.0			1.4	1.4				
13	Nonmonotone	Last tercile WBC	80.0	80.0	94	93	1.7	1.8				
14	Nonmonotone*	Last tercile WBC	80.0	80.0	91	91	1.4	1.4				

* Probability 1 of return to study once a visit is missed.

NOTE: $\hat{M}IP3$ represents the minimum of the estimated probabilities $\hat{\pi}_{i3}(\hat{\alpha})$ of remaining on study among subjects with $R_{i3} = 1$. $\hat{M}EP3$ represents the median of $\hat{\pi}_{i3}(\hat{\alpha})$ among subjects with $R_{i3} = 1$.

Downloaded by [Virginia Tech Libraries] at 04:07 17 July 2013

proposed in this article when nonresponse depends on past V_{it} history. The simulations were conducted so as to roughly approximate the AIDS data of Section 4. For simplicity, we simulated data from a single treatment arm (i.e., $X_i = 1$) and then computed the weighted estimator [Eq. (12)] of the treatment arm–specific mean CD4 count based on a number of selection models $\bar{\lambda}_{it}(\alpha)$ for $\bar{\lambda}_{it}$. For comparison, we first computed the sweep estimator and the optimal GEE estimator. We conducted two simulation experiments that differed in the choice of the correlation between CD4 counts and white blood cell counts. Each simulation experiment was based on 200 realizations, so that estimated coverage probabilities of true 95% confidence intervals will have a simulation accuracy of approximately ±3% due to Monte Carlo variability. In each realization, random-effects models were used to jointly generate CD4 counts Y_{it} and white blood cell counts V_{it} for 500 independent study subjects as follows: for $t = 0, 1, 2, 3$, $i = 1, \ldots, 500$, $Y_{it} = 200 - 40t + \sigma_{0i}(6 - t) + \varepsilon_{0it}$, $V_{it}^{1.33} = 3{,}000 - 100t + \sigma_{1i}(10 - t) + \varepsilon_{1it}$, where the random effects $(\sigma_{0i}, \sigma_{1i})$ were bivariate normal with mean zero, squared correlation coefficient ρ^2 either .81 or .36 depending on the experiment, and variances $(4.5^2, 100^2)$. The pure measurement errors ε_{1it} and ε_{0it} were generated independently for each subject i and each time t as follows: $\varepsilon_{1it} \sim N(0, 200^2)$ for all t, $\varepsilon_{0i0} \sim N(0, 40^2)$, $\varepsilon_{0i1} \sim N(0, 35^2)$, $\varepsilon_{0i2} \sim N(0, 25^2)$, and $\varepsilon_{0i3} \sim N(0, 10^2)$. These parameter values captured the following qualitative aspects of the AIDS data. The between-subject variance in the true CD4 counts, $(6 - t)^2\text{var}(\sigma_{0i})$, and the within-subject CD4 pure measurement error variance, $\text{var}(\varepsilon_{0it})$, both decrease with time t, reflecting the fact that as time increases, CD4 counts approach zero with declining marginal variance. In contrast, the time trends in the total white blood cell count and its variance are less marked. We used the power 1.33 in $V_{it}^{1.33}$ to capture the fact that CD4 count might not have a simple linear regression on white blood cell count.

In each simulation experiment, for each t, $\bar{\lambda}_{it}$ depended only on the population tercile of $V_{i(t-1)}$ with the conditional probability $\bar{\lambda}_{it}$ of remaining on study at t being .9, .75, and .5 for the highest, mid, and lowest terciles of $V_{i(t-1)}$. The data-generating and analysis programs were written in Borland's C^{++} Version 3.0 using the built in pseudo–random number generator and were implemented on a Northgate 386 personal computer. Results are given in Table 2. We report only results for time $t = 3$, because the results for $t = 1$ and $t = 2$ were similar, except that the biases were somewhat smaller. Examining rows 1–3 of Table 2, we observe that the sweep, optimal GEE, and sample average estimators are all biased upwards, with the bias of the sample average estimator being the greatest. As expected, because missingness depends on white blood cell count, the bias increases with increasing correlation ρ between white blood cell and CD4 counts. In row 4 we used a weighted estimator based on a separate logistic model $\bar{\lambda}_{it}(\alpha)$ at each occasion t, $t = 1, 2, 3$, with occasion-specific parameter vector α_t that included as covariates all past CD4 counts (linear CD4) so that $\alpha^T = (\alpha_1^T, \alpha_2^T, \alpha_3^T)$. This estimator performed in terms of bias and variance quite similarly to the GEE and sweep estimators. The poor coverage rates of the nominal confidence intervals in rows 1, 3, and 4 are due to the bias of $\hat{\beta}_3$ and not to poor variance estimation. In row 5 we observed that the weighted estimator based on a correctly specified logistic model for $\bar{\lambda}_{it}$ that, separately at each t, uses as a covariate an indicator for the tercile of white blood cell count at $t - 1$ was, as expected, unbiased. The weighted estimators in rows 5, 6, and 8 represent estimators based on a sequence of correctly specified nested models for $\bar{\lambda}_{it}$. As predicted by our theory, all are unbiased and the corresponding Monte Carlo variances of $\hat{\beta}$ are nonincreasing. As discussed by Robins et al. (1994, p. 897), at smaller sample sizes, our asymptotic results may not apply. As a result, the Monte Carlo variance of $\hat{\beta}$ based on a sequence of nested models may no longer be non-increasing, and the

actual coverage rates of Wald-type confidence intervals for β_0 based on the variance estimator in Theorem 1 may be less than nominal.

The interesting results in row 7 demonstrate the importance of accurately modeling the response probabilities $\bar{\lambda}_{it}$. The logistic model for $\bar{\lambda}_{it}$ in row 7 differs from that in row 6 in that white blood cell count at $t-1$ was included as a linear predictor rather than being transformed to a tercile indicator. Thus the model for $\bar{\lambda}_{it}$ in row 7 is misspecified. The bias of $\hat{\beta}$ in row 7 is greater in magnitude and opposite in direction to that in row 4. The reason for this can be understood by examining the $\hat{\text{M}}$IP3 column in Table 2. $\hat{\text{M}}$IP3 is the minimum of the estimated probability $\bar{\pi}_{i3}(\hat{\alpha})$ of remaining on study among subjects with $R_{i3}=1$. Hence $(\hat{\text{M}}\text{IP3})^{-1}$ is the maximum weight contributed by any of the 500 subjects to the estimator (12). Now by examining rows 5, 6, and 8, we see that for any correctly specified nonresponse model, $(\hat{\text{M}}\text{IP3})^{-1}$ is never greater that 10 and is close to the true value $[(.5)^3]^{-1}$ of $\{\min_i \bar{\pi}_{i3}(\alpha_0)\}^{-1}$. But for the misspecified model for $\bar{\lambda}_{it}$ of row 7, we obtained $(\hat{\text{M}}\text{IP3})^{-1}=50$, because subjects with extremely low white blood cell counts are incorrectly predicted to have only a 2% chance of surviving on study to t_3. Thus subjects with extremely low white blood cell counts are given five times the weight in the analysis in row 7 as they properly should. Hence when there is significant positive correlation between CD4 counts and white blood cell counts, subjects with low CD4 counts will be given inappropriately high weights, and considerable downward bias in $\hat{\beta}_3$ will occur. The message of this example is that when $\hat{\text{M}}$IP3 is very small (in comparison with the median estimated probability of remaining on study $\hat{\text{M}}$EP3), subjects i with the smallest values of $\bar{\pi}_{i3}(\hat{\alpha})$ will have a large influence on the analysis. Therefore, it is necessary to check carefully, using the model-checking techniques of Landwehr, Pregibon, and Shoemaker (1984), that the assignment of very low values of $\bar{\pi}_{i3}(\hat{\alpha})$ is appropriate (in that $\bar{\pi}_{i3}(\alpha_0)$ is truly small) and not, as in row 7, due to model misspecification. For pedagogic reasons, we purposely selected our simulation parameters to create an extreme situation in which model misspecification can result in $\hat{\text{M}}$IP3 much smaller than $\min_i\{\bar{\pi}_{i3}(\alpha_0)\}$. In contrast, in the analysis of the actual AIDS data reported in Section 4, the minimum value of $\hat{\text{M}}$IP3 taken over all the models used for $\bar{\lambda}_{it}$ and both treatment arms was only .23. The misspecification of the model for $\bar{\lambda}_{it}$ in Row 7 also leads to excessive variability in $\hat{\beta}_3$, because the magnitude of $\hat{\beta}_3$ in any realization depends strongly on the value of Y_{i3} among subjects with weights near 50.

The results in row 9 of Table 2 are based on the extended sweep estimator. By definition, when $X_i=1$, the extended sweep estimator is the MLE of $E(Y_i)$ under the assumptions (a) $Y_i \equiv \bar{Y}_{i(T+1)}$ and $V_i \equiv \bar{V}_{i(T+1)}$ are jointly multivariate normal with unrestricted mean and covariance matrix and (b) the data are missing at random; that is, Equation (3) is true (see Little and Rubin 1987 and Appendix B herein). The extended sweep estimator uses data on white blood cell count, V_{it}. In contrast, the usual sweep estimator previously reported in row 3 of Table 2 and in Table 1 is the MLE of $E(Y_i)$ when (a) Y_i is multivariate normal, (b) Equation (3) is true with W_{it} redefined to be $\bar{Y}_{it}$, and (c) the data on V_{it} are ignored. The extended sweep estimator still has a bias of $-.4$ when $\rho^2=.81$. Nonetheless, the bias is much less than that of the usual sweep estimator of row 3. Theorem B.2 in the Appendix implies that even if the data are not normally distributed, when (2) holds and $X_i=1$, the extended sweep estimator would be asymptotically unbiased if W_{it} had a linear regression on $\bar{W}_{it}$; that is, for all t,

$$E(W_{it}\mid \bar{W}_{it}, R_{it}=1)=b_t\bar{W}_{it}. \tag{13}$$

Because $V_{it}^{1,33}$ and Y_{it} were simulated as jointly normal, this implies that, as is evident from row 10, the extended sweep operator is unbiased when we replace V_{it} by $V_{it}^{1,33}$. Furthermore, the variance in row 10 is less than that of our most efficient weighted estimator in row 8, because when nonresponse depends on V_{it}, imposing the assumption (13) supplies additional information about $E(Y_{i3})$.

Rows 11 and 12 consider the extended optimal GEE estimator. When $X_i=1$, the extended optimal GEE estimator (a) solves $\Sigma_i C_i^{*-1}\varepsilon_i^*(\beta)=0$, with $\varepsilon_i^*(\beta)=((Y_i^*-\beta_y)^T, (V_i^*-\beta_v)^T)$ redefined to be the vector of observed residuals for both Y_i and V_i where $\beta^T=(\beta_y^T, \beta_v^T)$, and (b) uses as a working covariance matrix C_i^*, the empirical covariance matrix of the observed residuals $\varepsilon_i^*(\beta)$ based on all "available cases." It is evident from row 12 that, as predicted by theory, the extended optimal GEE estimator, in contrast to the extended sweep estimator, remains biased even when based on $V_{it}^{1,33}$.

The results in rows 9 and 10 suggest that if (13) is true (possibly after transformation of the V_{it} variables), then the extended sweep operator may be an alternative to a weighted estimator when estimating treatment arm–specific CD4 count means, because the extended sweep estimator may be more efficient and because it does not require a correctly specified model for the probabilities $\bar{\lambda}_{it}$ of response. The extended sweep estimator does have drawbacks, however. First, the conditional mean of W_{it} given $\bar{W}_{it}$ will often be a nonlinear function of $\bar{W}_{it}$ (particularly if W_{it} has Bernoulli components), in which case (13) will be false and therefore the extended sweep estimator may be inconsistent when the data are missing at random but not completely at random; that is, Equation (7) is false. In contrast, although consistency of the weighted estimators requires (8) to hold, there are no restrictions on the functional form of the model $\bar{\lambda}_{it}(\alpha)$. A second drawback to the extended sweep operator is that when X_i contains pretreatment variables such as age, the extended sweep estimator of $E(Y_{it}\mid X_i)$ described in Appendix B will generally be consistent only if both (13) holds and (1) is true with $g_t(X_i,\beta)=\beta_t X_i$, where $\beta=(\beta_1,\ldots\beta_T)^T$. But $g_t(X_i,\beta)$ will often be nonlinear, especially when Y_{it} is Bernoulli. In Appendix B we briefly consider a possible nonlinear generalization of the sweep estimator—the iterated conditional expectations estimator.

G-Computation Algorithm Estimators, the Intent-to-Treat Paradigm, and Nonparametric Estimation

When Equation (5b) holds, yet another approach to the estimation of $E[Y_{it}\mid X_i]=x, x=0,1$ is to use the likelihood-

Downloaded by [Virginia Tech Libraries] at 04:07 17 July 2013

Downloaded by [Virginia Tech Libraries] at 04:07 17 July 2013

based parametric G-computation algorithm estimator (Robins 1987b, app. 1; 1989, sec. 5), in which one specifies parametric models $f(w_j|\bar{w}_j, R_{ij} = 1, X_i = x; \omega), j = 0, \ldots, t$, and then evaluates, by Monte Carlo methods if necessary, the G-computation algorithm multiple integral formula (5a) with $f(w_j|\bar{w}_j, R_{ij} = 1, X_i = x; \hat{\omega})$ replacing $f(w_j|\bar{w}_j, R_{ij} = 1, X_i = x), j = 0, \ldots, t$, where $\hat{\omega}$ is the MLE of ω. The generalized sweep estimator of $E[Y_{it}|X_i = x]$ is a special case of the G-computation algorithm in which the models $f(w_j|\bar{w}_j, R_{ij} = 1, X_i = x; \omega)$ specify that (13) holds with residuals $W_{it} - b_t\bar{W}_{it}$ normally distributed with mean zero and variance matrix σ_t^2 and $\omega = (b_t, \sigma_t^2; t = 1, \ldots, T)$. If, in addition to Equation (2a), Equation (7) is true so that the data are missing completely at random (i.e., Eq. (6) holds), then the usual and extended sweep, the "optimal" GEE, and weighted estimators of the difference in treatment arm–specific means are guaranteed to be consistent (provided of course that the treatment arm indicator is included as a covariate in the models $\lambda_{it}(\alpha)$). In contrast, for most, $f(w_j|\bar{w}_j, R_{ij} = 1, X_i; \omega)$, for which either (13) is false or $W_{it} - b_t\bar{W}_{it}$ is non-normal, the parametric G-computation algorithm estimator will be inconsistent if the models are misspecified. We believe many investigators would be reluctant to estimate the treatment effect $E(Y_{it}|X_i = 1) - E(Y_{it}|X_i = 0)$ using a method (designed to help correct for dependent censoring and nonrandom noncompliance) that can be inconsistent in the presence of either (a) independent censoring and random noncompliance (i.e., in the presence of data missing completely at random) or (b) complete compliance with censoring absent. An estimator, such as the parametric G-computation algorithm estimator, that can be inconsistent under both (a) and (b) violates the spirit of the intent-to-treat paradigm for the analysis of randomized trials in which (a) and/or (b) is/are true. But, if one were to estimate the conditional densities $f(w_j|\bar{w}_j, R_{ij} = 1, X_i)$ completely nonparametrically (e.g., by using kernel estimators of appropriately chosen bandwidth when w_j has one or more continuous components), then, under independent censoring (i.e., when (2a) and (7) are true), the resulting "nonparametric" Monte Carlo G-computation algorithm estimator would be guaranteed to be asymptotically normal and unbiased for $E(Y_{it}|X_i = 1) - E(Y_{it}|X_i = 0)$ and would be efficient in the sense of Begun et al. (1983) with asymptotic variance given by the lower bound provided in Lemma 1. But as discussed by Huber (1985), if the dimension of $\bar{W}_{ij}$ is large, then, due to the "curse of dimensionality," nonparametric estimation of the densities $f(w_j|\bar{w}_j, R_{ij} = 1, X_i)$ will not be practical and the nonparametric G-computation algorithm estimator cannot be used.

When either (7) is false or X_i has continuous components, we cannot be certain that our models $\bar{\lambda}_{it}(\alpha)$ are correct. Thus, as an alternative to using $\bar{\lambda}_{it}(\hat{\alpha})$ in the estimating Equation (10), we could have replaced $\bar{\lambda}_{it}(\hat{\alpha})$ by a completely nonparametric multivariate kernel regression estimator of $\bar{\lambda}_{it}$ to protect against misspecification bias like that found in rows 4 and 7 of Table 2. Although detailed consideration of the large-sample properties of the resulting "nonparametric" weighted estimator is beyond the scope of this article, it can be shown that when (5b) is true, under sufficient smoothness and regularity assumptions and with the bandwidth properly chosen, the nonparametric weighted estimator will be asymptotically normal with an asymptotic variance that attains the efficiency bound described in Lemma 1 and will be asymptotically equivalent to the nonparametric G-computation algorithm estimator. Of course, due to the "curse of dimensionality," when the dimension of $\bar{W}_{it}$ is large, some model for $\bar{\lambda}_{it}$ will be required in practice, although the model need not be parametric. For example, one could use a generalized additive model.

6. EXTENSION TO ARBITRARY MISSING DATA PATTERNS

The assumption that the missing data pattern is monotone is restrictive because, in practice, subjects typically may miss several visits and then show up for a subsequent visit. To allow for arbitrary missing data patterns, set $R_{it}^* = R_{it}$ so that $R_i^* \equiv (R_{i1}^*, \ldots, R_{iT}^*)^T$ is the vector of missing visit indicators. We now allow R_i^* to take on any of its 2^T possible realizations $r = (r_1, \ldots, r_T)^T$, where r is a T-vector of 1s and zeros. Now *redefine* R_{it} so that $R_{it} = 1$ if $R_{i1}^* = R_{i2}^* = \cdots = R_{it}^* = 1$ and $R_{it} = 0$ otherwise. Hence R_{it} is now an artificial "censoring indicator" that is zero once a subject fails to return at any occasion, t', for $t' \le t$. Note that $R_{iT} = 1 \Leftrightarrow R_i^* = \mathbf{1}$, where $\mathbf{1}$ is the T-vector of 1s. One approach to analyzing nonmonotone missing data is to use the R_{it} to "artificially" treat the data as monotone and use analytic methods appropriate for monotone data (Little and Rubin 1987). Specifically, if (1), (2), and (8) are true, then the "artificial monotone" estimator $\hat{\beta}$ solving (10) remains asymptotically normal and unbiased for β_0, where we define an estimator to be "artificially monotone" if it depends on the data only through the data M_i available on subject i until the first missed visit. That is, $M_i = (\bar{W}_{i(T+1)}, R_{i1}, \ldots, R_{iT})$ if $R_{iT} = 1$ and $M_i = (\bar{W}_{it}, R_{i1}, \ldots, R_{iT})$ if $R_{i(t-1)} - R_{it} = 1$. [It is of interest to note that when some dropouts later return to the study, the missingness process in the semiparametric model characterized by (1), (2), (8) is nonignorable (i.e., that data are no longer missing at random). Even so, $\hat{\beta}$ remains consistent for β_0.] Because this approach disregards the subsequent data obtained on a given subject once a visit is missed, one might be concerned that it will be quite inefficient if many subjects miss early visits but later return. Nevertheless, when at least some of the dropouts at each visit t do not later return, Robins and Rotnitzky (1995) show that unless one is willing to impose additional assumptions concerning the missing data mechanism over and above (2), no information concerning β_0 can be extracted from the additional visits of subjects who have once missed a visit. But suppose now that we are willing to impose the additional assumption

$$P(R_{it}^* = 1|\bar{R}_{it}^*, \bar{W}_{it}^*, \bar{W}_{i(T+1)}) = P(R_{it}^* = 1|\bar{R}_{it}^*, \bar{W}_{it}^*), \quad (14)$$

where $\bar{R}_{it}^* = (R_{i1}^*, \ldots, R_{i(t-1)}^*)$ is a subject's past nonresponse history and $\bar{W}_{it}^* = (W_{i0}^*, W_{i1}^*, \ldots, W_{i(t-1)}^*)$, here $W_{i0}^* = W_{i0}$, $W_{it}^* = R_{it}^* W_{it}$ is a subject's observed past W_{it} history.

Downloaded by [Virginia Tech Libraries] at 04:07 17 July 2013

Equation (14) states that the probability of being observed at t, given the observed past $(\bar{R}^*_{it}, \bar{W}^*_{it})$ through $t-1$, does not depend on the unobserved past or on the future, and it is stronger than Equation (2a). Further, Equation (14) implies that the data are missing at random (MAR) in the sense of Rubin (1976), although missing at random does not imply (14). In fact, in contrast to MAR, (14) is an identifiable restriction and thus is subject to empirical test, although the construction of such a test is beyond the scope of this article. Mark and Gail (1994) have considered a missing data mechanism that is mathematically equivalent to (14). We now show how under assumption (14), we can construct estimators $\tilde{\beta}$ that are more efficient than estimators $\hat{\beta}$ solving (10). We motivate the estimator $\tilde{\beta}$ by first defining an estimator $\tilde{\beta}^\dagger$ that is not available for data analysis because it depends on unknown population quantities. Redefine $\bar{\lambda}_{it} = P(R^*_{it} = 1 | \bar{R}^*_{it}, \bar{W}^*_{it})$ and let $\bar{\lambda}_{it}(\alpha)$ be a correctly specified model for $\bar{\lambda}_{it}$; that is, $\bar{\lambda}_{it}(\alpha) = \lambda_t(\alpha, \bar{R}^*_{it}, \bar{W}^*_{it})$ is now a known function of α, $\bar{R}^*_{it}$, $\bar{W}^*_{it}$ taking values in (0, 1] satisfying Equation (8). Let $\hat{\alpha}$ solve $\Sigma_i S_{\alpha,i}(\alpha) = 0$, where now $S_{\alpha,i}(\alpha) = \partial \log\{\Pi_t[\bar{\lambda}_{it}(\alpha)]^{R^*_{it}}[1 - \bar{\lambda}_{it}(\alpha)]^{1-R^*_{it}}\}/\partial\alpha = \sum_{t=1}^{T}[R^*_{it} - \bar{\lambda}_{it}(\alpha)]\partial \text{ logit } \bar{\lambda}_{it}(\alpha)/\partial\alpha$.

To define $\tilde{\beta}^\dagger$, set $\bar{W}^*_{it}(r) = (W_{i0}, r_1 W_{i1}, \ldots, r_{t-1} W_{i(t+1)})$ and $\bar{\pi}_i(r, \alpha) \equiv \prod_{t=1}^{T} \lambda_t(\alpha, \bar{r}_t, \bar{W}^*_{it}(r))^{r_t}[1 - \lambda_t(\alpha, \bar{r}_t, \bar{W}^*_{it}(r))]^{1-r_t}$. Note that $\bar{W}^*_{it}(r)$ is the observed W_{it} history through $t-1$ of a subject i with nonresponse history $R^*_i = r$. In particular, note that $\bar{W}^*_{it}(R^*_i) = \bar{W}^*_{it}$. When (14) and (8) hold, $\bar{\pi}_i(r, \alpha_0) = P(R^*_i = r | \bar{W}_{i(T+1)})$. That is, $\bar{\pi}_i(r, \alpha_0)$ is the probability that a subject i with a given $\bar{W}_{i(T+1)}$ has a nonresponse trajectory given by the vector r. Note that $\bar{\pi}_i(\mathbf{1}, \alpha)$ equals $\bar{\pi}_{iT}(\alpha)$. Let $\phi_i = \{\phi_{i,r}; r \neq \mathbf{1}\}$, where $\phi_{i,r}$ is, for each $r \neq \mathbf{1}$, a known ν-dimensional function of $\bar{W}^*_{i(T+1)}(r)$ selected by the investigator. For example, in the simulation results reported later, ν is 2, V_{it} is univariate, and $\phi_{i,r}$ is the vector $(V_{it(r)}, Y_{it(r)})^T$ obtained at the last available visit for a subject with trajectory r; that is, $t(r)$ is the largest value of t such that $r_t = 1$. Define $A_i(\phi, \hat{\alpha}) = R_{iT}\{\bar{\pi}_{iT}(\hat{\alpha})\}^{-1}\Sigma_{r\neq\mathbf{1}}\bar{\pi}_i(r, \hat{\alpha})\phi_{i,r} - \Sigma_{r\neq\mathbf{1}} I(R^*_i = r)\phi_{i,r}$. Note that $A_i(\phi, \hat{\alpha})$ is a function of $\hat{\alpha}$, the observed data, and the functions $\phi_{i,r}$ chosen by the investigator, and thus can be evaluated for each subject i. Further, $A_i(\phi, \alpha_0)$ is not a function of β and, as is shown in Appendix A, has mean zero when (14) and (8) are true. Let $\tilde{\beta}^\dagger$ solve

$$0 = U^\dagger(\beta, \phi, \hat{\alpha}),$$

where $U^\dagger(\beta, \phi, \hat{\alpha}) = n^{-1/2}\Sigma_i U_i^\dagger(\beta, \phi, \hat{\alpha})$ with $U_i^\dagger(\beta, \phi, \hat{\alpha}) = U_i(\beta, \hat{\alpha}) - \theta A_i(\phi, \hat{\alpha})$ and $\theta = E[\text{Resid}\{U_i, S_{\alpha,i}\} \text{Resid}\{A_i(\phi, \alpha_0), S_{\alpha,i}\}^T][\text{var}[\text{Resid}\{A_i(\phi, \alpha_0), S_{\alpha,i}\}]]^{-1}$.

The solution $\tilde{\beta}^\dagger$ is consistent for β_0 when (1), (2b), (8), and (14) are true. The key to the consistency of $\tilde{\beta}^\dagger$ is the fact that under (8) and (14), $E[A_i(\phi, \alpha_0)] = 0$. The choice of θ may seem arbitrary. But θ, as defined previously, was chosen to guarantee that $\tilde{\beta}^\dagger$ is at least as efficient as $\hat{\beta}$ solving (10). In fact, $\tilde{\beta}^\dagger$ will be strictly more efficient than $\hat{\beta}$ whenever $\theta \neq 0$. This occurs because, as is shown in Appendix A, for this choice of θ,

$$n^{1/2}(\tilde{\beta}^\dagger - \beta_0) = -E\{\partial U_i^\dagger(\beta_0, \phi, \alpha_0)/\partial\beta^T\}^{-1} n^{-1/2} \times \Sigma_i \text{Resid}(U_i, B_i) + o_p(1), \quad (15)$$

where $B_i^T = (A_i(\phi, \alpha_0)^T, S_{\alpha,i}^T)$ and, further, $E[\partial U_i^\dagger(\beta_0, \phi, \alpha_0)/\partial\beta^T] = E[\partial U_i(\beta_0, \alpha_0)/\partial\beta^T] = \Gamma$ because $A_i(\phi, \alpha_0)$ does not depend on β. Thus by the usual sandwich formula, $n^{1/2}(\tilde{\beta}^\dagger - \beta_0)$ has asymptotic variance $\Gamma^{-1}\text{var}\{\text{Resid}(U_i, B_i)\}\Gamma^{-1T}$. As discussed in Section 2, the asymptotic variance of $n^{1/2}(\hat{\beta} - \beta_0)$ is $\Gamma^{-1}\text{var}\{\text{Resid}(U_i, S_{\alpha,i})\}\Gamma^{-1T}$. Thus $\hat{\beta}$ is never more efficient than $\tilde{\beta}^\dagger$ because, $S_{\alpha,i}$ being a subvector of B_i implies that, by standard least squares theory, $\text{var}\{\text{Resid}(U_i, B_i)\} \leq \text{var}\{\text{Resid}(U_i, S_{\alpha,i})\}$. Furthermore, we show in Appendix A that θ equals the coefficient of $A_i(\phi, \alpha_0)$ in the population regression of U_i on B_i; thus unless $\theta = 0$, the last inequality is strict and $\tilde{\beta}^\dagger$ will be more efficient than $\hat{\beta}$. An analogous argument can be used to show that if we increase the dimension ν of $\phi_{i,r}$ by adding additional components such as $Y^2_{it(r)}$ to $(Y_{it(r)}, V_{it(r)})$, the asymptotic variance of $\tilde{\beta}^\dagger$ will never increase and will usually decrease.

Because θ is an unknown population quantity, $\tilde{\beta}^\dagger$ is not available for data analysis. But, as is stated in the following Theorem, proved in Appendix A, $\tilde{\beta}$ solving

$$0 = \hat{U}^\dagger(\beta, \phi, \hat{\alpha}) \quad (16)$$

with $\hat{U}^\dagger(\beta, \phi, \hat{\alpha}) = n^{-1/2}\Sigma_i U_i(\beta, \hat{\alpha}) - \hat{\theta}A_i(\phi, \hat{\alpha})$ will have the same limiting distribution as $\tilde{\beta}^\dagger$ where $\hat{\theta}$ is the $n^{1/2}$-consistent estimator of θ given by $\hat{\theta} = \hat{\theta}_1\hat{\theta}_2^{-1}$, $\hat{\theta}_1 = n^{-1}\Sigma_i \hat{\text{Resid}}[U_i(\hat{\beta}, \hat{\alpha}), S_{\alpha,i}(\hat{\alpha})]\hat{\text{Resid}}[A_i(\phi, \hat{\alpha}), S_{\alpha,i}(\hat{\alpha})]^T$, and $\hat{\theta}_2 = n^{-1}\Sigma_i\{\hat{\text{Resid}}[A_i(\phi, \hat{\alpha}), S_{\alpha,i}(\hat{\alpha})]\}^{\otimes 2}$.

Theorem 2. Subject to the regularity conditions in the Appendix A, if (1), (8), (2b), and (14) are true, then, with probability approaching 1, (a) there exists a unique solution $\tilde{\beta}$ to (16) and (b) $n^{1/2}(\tilde{\beta} - \beta_0)$ is asymptotically normal with mean zero and asymptotic variance $\Gamma^{-1}\text{var}\{\text{Resid}(U_i, B_i)\}\Gamma^{-1T}$, which can be consistently estimated by $\hat{\Gamma}^{-1}[n^{-1}\Sigma_i\{\hat{\text{Resid}}[U_i(\hat{\beta}, \hat{\alpha}), \{A_i(\phi, \hat{\alpha})^T, S_{\alpha,i}(\hat{\alpha})^T\}^T]\}^{\otimes 2}]\hat{\Gamma}^{-1T}$.

In the special case in which X_i is a dichotomous variable,

$$\tilde{\beta}_{x,t} = \frac{\sum_{i=1}^{n} I(X_i = x)\bar{\pi}_{it}(\hat{\alpha})^{-1}R_{it}Y_{it} - \hat{\theta}A_i(\phi, \hat{\alpha})}{\sum_{i=1}^{n} I(X_i = x)\bar{\pi}_{it}(\hat{\alpha})^{-1}R_{it}}$$

is the estimate of $E(Y_{it}|X_i = x)$ that solves (16). Row 13 of Table 1 studies the performance of $\tilde{\beta}_{1,3}$ in the simulation experiment of Section 5, modified so that Equation (14) was true and $\bar{\lambda}_{it}$ was .9, .75, and .5 for the 1st, 2nd, and 3rd terciles of $V_{i(t-1)}$ if $R^*_{i(t-1)} = 1$; .8, .65, and.4 for the 1st, 2nd, and 3rd terciles of $V_{i(t-2)}$ if $R^*_{i(t-1)} = 0$ and $R^*_{i(t-2)} = 1$; and zero if $R^*_{i(t-1)} = R^*_{i(t-2)} = 0$, so that no subject returns after missing two visits. Write $\alpha^T = (\alpha^{(1)T}, \alpha^{(2)T})$. Then in estimating $\bar{\lambda}_{it}$, (a) for $R_{i(t-1)} = 1$, $\bar{\lambda}_{it}(\alpha)$ was the nonresponse model with parameter vector $\alpha^{(1)}$ of row 5 of Table 2; (b) when $t = 3$, $R_{i(t-1)} = 0$, and $R^*_{i(t-1)} = 1$ (i.e., $R^*_{i1} = 0$ and $R^*_{i2} = 1$), $\bar{\lambda}_{it}(\alpha)$ was a logistic model with parameter vector $\alpha_1^{(2)}$ in terciles of $V_{i(t-1)}$; (c) when $R^*_{i(t-1)} = 0$ and $R^*_{i(t-2)} = 1$, $\bar{\lambda}_{it}(\alpha)$ was a logistic model in terciles of $V_{i(t-2)}$ with parameter vectors $\alpha_2^{(2)}$ at $t = 2$ and $\alpha_3^{(2)}$ at $t = 3$; and (d) $\bar{\lambda}_{it}(\alpha)$ was set to 0 if $R^*_{i(t-1)} = R^*_{i(t-2)} = 0$. Finally, as described earlier, $\phi_{i,r}$ was the vector $(V_{it(r)}, Y_{it(r)})^T$. As predicted by theory, $\tilde{\beta}_{1,3}$ was never less efficient than the estimator $\hat{\beta}_{1.3}$ of row 5 and was slightly more efficient when $\rho^2 = .81$.

To demonstrate the large theoretical gains in efficiency available by using $\tilde{\beta}$, we studied, in row 14 of Table 1, $\tilde{\beta}_{1,3}$ in the rather unlikely setting in which once a subject had

Downloaded by [Virginia Tech Libraries] at 04:07 17 July 2013

missed a visit, the subject always returned at the next occasion and missed no more visits. That is, we set $\bar{\lambda}_{it} = 1$ if $R^*_{it'} = 0$ for any $t' \leq t$. As before, for a subject who had never missed a visit prior to t (i.e., $R_{i(t-1)} = 1$), we set $\bar{\lambda}_{it}$ to .9, .75, and .5 for the 1st, 2nd, and 3rd terciles of $V_{i(t-1)}$. Under this design, the only subjects who had Y_{i3} missing are subjects who missed their first visit at $t = 3$. In analyzing the data we used the same model $\bar{\lambda}_{it}(\alpha)$ and $\phi_{i,r}$ as in row 13. Reading from row 14, we see that $\tilde{\beta}$ provides striking efficiency gains over $\hat{\beta}$ of row 5 when all subjects who left prior to $t = 3$ later returned. In row 14 the actual coverage of the nominal 95% confidence intervals was only 91%. This undercoverage is no longer present at a sample size of 1,000 (data not shown).

Because the class of estimators $\tilde{\beta}$ was proposed on an ad hoc basis, one might be concerned that such estimators fail to extract all the additional information concerning β_0 that follows from imposing assumption (14). But Robins and Rotnitzky (1995) show that there exists an estimator $\tilde{\beta}_{op}$ in our class that attains the semiparametric variance bound for the semiparametric model characterized by (1), (2a), (8), and (14). Because, as noted earlier, $\hat{\beta}_{op}$ is semiparametric efficient in the less restrictive model that assumes only (1), (2a), (2b), and (8) are true, $\tilde{\beta}_{op}$ will be more efficient than $\hat{\beta}_{op}$, and the ratio of the asymptotic variances of $\hat{\beta}_{op}$ to that of $\tilde{\beta}_{op}$ is a measure of the improvement in efficiency that can be obtained when (14) is (correctly) assumed to be true. From a substantive viewpoint, we believe that it is much more likely that assumption (2a) is approximately true than that assumption (14) is approximately true. Thus in practice we would recommend some caution in using the estimator $\tilde{\beta}$ proposed in this section.

Remark. Because, given the redefinition of $\bar{\lambda}_{it}(\alpha)$, $\hat{\alpha}$ will now depend on data obtained after a subject has missed a visit, the solution $\hat{\beta}$ to (10) that uses $\bar{\lambda}_{it}(\alpha)$ as redefined need no longer be an "artificial monotone" estimator. But if, as in the simulation experiment described in this section, $\alpha^T = (\alpha^{(1)T}, \alpha^{(2)T})$, $\alpha^{(1)}$ and $\alpha^{(2)}$ are variation independent, and $\bar{\lambda}_{it}(\alpha)$ depends only on $\alpha^{(1)}$ if $R_{i(t-1)} = 1$ and $\bar{\lambda}_{it}(\alpha)$ depends only on $\alpha^{(2)}$ if $R_{i(t-1)} = 0$, then $\hat{\beta}$ will be an "artificial monotone" estimator, because $\bar{\pi}_{it}(\alpha)$ depends only on $\alpha^{(1)}$ and $\hat{\alpha}^{(1)}$ is obtained by maximizing Equation (9), which depends only on the artificial monotone data M_i, $i = 1, \ldots, n$.

7. SUMMARY AND FURTHER CONSIDERATIONS

We have proposed asymptotically unbiased weighted estimators of the parameters β_0 of the regression of a multivariate response variable Y_i on a set of explanatory variables X_i in the presence of missing data. Our methods extend the generalized estimating equations proposed by Liang and Zeger to settings where nonresponse can depend both on past outcomes and covariate history. As with the GEE, with our methods we do not need to specify the complete data likelihood. In contrast to a likelihood approach, our techniques require correct specification of the probability of nonresponse given the observed past $\bar{W}_{it}$. Thus our methods will be particularly appealing for analyzing certain non-Gaussian data, because some complete data likelihoods satisfying the restriction (1) can be cumbersome when X_i and/or Y_i are multivariate, whereas, in contrast, models for the nonresponse probabilities $\bar{\lambda}_{it}(\alpha)$ are easy to specify. Further, even when X_i is the dichotomous treatment arm indicator of a randomized trial, parametric likelihood-based estimators (i.e., the parametric G-computation algorithm estimator) can, in contrast to our weighted estimator, be inconsistent for the difference in treatment arm–specific means due to model misspecification even when compliance is completely at random and censoring is independent.

The results of Theorems 1 and 2 hold, however, even if the explanatory variables X_i are not independent and identically distributed, provided that the vectors Y_i, $i = 1, \ldots, n$, are independently distributed conditional on the sequence X_i, $i = 1, \ldots, n$, because in that case $\{X_i: i = 1, \ldots, n\}$ will be ancillary for β_0. An example of a setting where this occurs would be the randomized trial of Section 2 if treatment assignment was based on a randomized block design.

When (as in the environmental tobacco smoke example of Section 2) data on X_{it} are not available when a subject misses occasion t (i.e., $R^*_{it} = 0$), one can compute $U_i(\beta, \alpha) = D_i(\beta)\Delta_i(\alpha)\varepsilon_i(\beta)$ from the observed data only if the tth column of $D_i(\beta)$ depends on $\bar{X}_{i(t+1)}$ but not on $(X_{i(t+1)}, \ldots, X_{iT})$. Subject to this restriction on the columns of $D_i(\beta)$, the results of the article remain true.

When the components X_{it} of X_i are time dependent and do not satisfy the definition of an ancillary external covariate process (i.e., $f[X_{it}|\bar{W}_{it}] = f[X_{it}|\bar{X}_{it}]$, $t = 1, \ldots, T$ is false), the parameter β_0 of model (1) will fail to have a causal interpretation even if model (1) is correctly specified and there are no missing data. (Note that $\bar{X}_{it}$ is a function of $\bar{W}_{it}$.) In this setting one can use the structural nested mean models described by Robins (1989, sec. 17, 1994a) to estimate the causal effect of past $\bar{X}_{it}$ history on the mean of Y_{it}.

Suppose that the missing data pattern is monotone but that we no longer impose the nonidentifiable assumption Equation (5b) because we believe that it may be false. Then Equation (2a), the missing at random assumption (3), and the G-computation algorithm formula (5a) will also be false, and $E[Y_{it}|X_i]$ is no longer identified. In this setting of nonignorable missingness, Little and Rubin (1987) suggested that one estimate the parameter β_0 of Equation (4) under varying assumptions about the missing data mechanism. To do so, redefine $\bar{\lambda}_{it} = P[R_{it} = 1 | R_{i(t-1)} = 1, \bar{W}_{it}, Y_i]$ and $\bar{\pi}_{it} = \prod_{j=1}^{t} \bar{\lambda}_{ij}$. Then if $\bar{\lambda}_{it}$ and $\bar{\pi}_{it}$ are unrestricted they are no longer identifiable based on the observed data; however, the identities $E[\bar{\pi}_{it}^{-1} R_{it} Y_{it}|X_i] = E[Y_{it}|X_i]$ and $E[U_i(\beta_0)|X_i] = 0$ remain true. Here $U_i(\beta) \equiv D_i(\beta)\Delta_i\varepsilon_i(\beta)$ with $\Delta_i \equiv \text{diag}\{\bar{\pi}_{it}^{-1} R_{it}\}$. Thus we can perform a sensitivity analysis by displaying the $\hat{\beta}$ solving $\Sigma_i U_i(\beta) = 0$ for various choices of the $\bar{\lambda}_{it}$, with each choice representing a different assumption concerning the nonidentifiable dependence of $\bar{\lambda}_{it}$ on Y_i. We prefer this sensitivity analysis to one in which we redefine $\bar{\lambda}_{it}$ to be $P[R_{it} = 1 | R_{i(t-1)} = 1, Y_i]$, because the dependence on the unobserved components of Y_i of $P[R_{it} = 1 | R_{i(t-1)} = 1, \bar{W}_{it}, Y_i]$ will typically be less than that of $P[R_{it} = 1 | R_{i(t-1)} = 1, Y_i]$ when $\bar{W}_{it}$ is both a strong predictor of dropout and is correlated with Y_i's unobserved components. More generally, suppose that missingness may be non-

monotone as well as non-ignorable, adopt the notation of Sec. 6, and now redefine $\bar{\lambda}_{it} = P[R_{it}^* = 1 | \bar{R}_{it}^*, \bar{W}_{i(T+1)}]$. Now, rather than simply perform a sensitivity analysis under various a priori choices for $\bar{\lambda}_{it}$, suppose, following Diggle and Kenward (1994), we choose to jointly estimate β_0 and $\bar{\lambda}_{it}$ from the data by (correctly) specifying a model $\bar{\lambda}_{it}(\alpha) = \lambda_t(\alpha, \bar{R}_{it}^*, \bar{W}_{i(T+1)})$ for the missingness process $\bar{\lambda}_{it}$, with $\lambda_t(\cdot, \cdot, \cdot)$ a known function and α a $q \times 1$ unknown parameter vector. To jointly estimate $\psi_0 = (\beta_0^T, \alpha_0^T)^T$ in defining $U_i(\beta, \alpha)$, let $D_i(\beta) = d(X_i, \beta)$ now be a $(p + q) \times T$ matrix and choose the dimension ν of $\phi_{i,r}$ to be $p + q$. Let $\tilde{\psi} = (\tilde{\beta}^T, \tilde{\alpha}^T)^T$ solve $\Sigma_i U_i(\beta, \alpha) - A_i(\phi, \alpha) = 0$. Robins and Rotnitzky (1995) prove that, in the semiparametric model characterized by (1), (2b) and the model $\bar{\lambda}_{it}(\alpha)$, when ψ_0 is identified, (a) $\tilde{\psi}$ is asymptotically normal and unbiased for ψ_0, and (b) if $\mathring{\psi} = (\mathring{\beta}^T, \mathring{\alpha}^T)^T$ is a regular, asymptotically linear estimator of ψ_0, then $\mathring{\psi}$ is asymptotically equivalent to $\tilde{\psi}$ for some choice of the functions $d(\cdot, \cdot)$ and ϕ.

Finally, consider again the hypothetical randomized trial of Section 2. If the large fraction of subjects, say 60%, initiates therapy with the nonrandomized treatment B early in the study, then the efficiency of any weighted estimator will be poor. In this setting, instead of regarding the subject as censored at time of initiating therapy with drug B, we might estimate the effect of the randomized treatment A on Y_{it} by using structural nested mean models to estimate the independent effect of therapy B on the mean of Y_{it} (Robins 1989, sec. 17; 1994a) and then subtracting off that effect from the observed mean of Y_{it}. Robins (1989, 1993a), Mark and Robins (1993ab), and Robins and Greenland (1994) carried out a similar program in the analysis of randomized trials with failure time outcomes using structural nested failure time models in the presence of nonrandom noncompliance and dependent censoring; in the same setting, Robins (1991, 1993b, 1995b), and Robins and Rotnitzky (1992), provided inverse-probability-of-censoring weighted estimators, analogous to those proposed in this article, that can be used to estimate the parameters of the Cox proportional hazards model a, median regression failure time model and the accelerated failure time model. Robins (1993b) compared these weighted estimates to the G-computation algorithm estimators of Lagakos (1977), Gray (1993), Finkelstein and Schoenfeld (1994) and Robins (1986), as well as to the generalized Buckley–James estimator described by Robins and Rotnitzky (1992) and the estimator of Malani (1992) and Fleming, Prentice, Pepe, and Glidden (1994). A wholly different approach to adjusting for dependent censoring in randomized trials when data on surrogates $\vec{V}_t$ are not obtained has been proposed for both failure time and repeated measures outcomes by Robins and Rotnitzky (1992, Appendix 4), Lin, Robins, and Wei (1995), and Robins (1994b).

APPENDIX A: PROOFS

We first establish the following lemma.

Lemma A.1. Given (2b), the G-computation algorithm formula (5a) is true if and only if $E[R_{it}Y_{it}/\bar{\pi}_{it}|X_i] = E[Y_{it}|X_i]$.

Proof. We will show that the right side of (5a) equals $E[R_{it}Y_{it}/\bar{\pi}_{it}|X_i]$. To do so, multiply within the integral in (5a) by $\bar{\pi}_i(\bar{w}_t)/\bar{\pi}_i(\bar{w}_t)$ with $\bar{\pi}_i(\bar{w}_t) \equiv \prod_{j=1}^{t} P(R_{ij} = 1 | \bar{w}_j, R_{i(j-1)} = 1, X_i)$ to obtain that the right side of (5a) equals $E\{E[Y_{it}/\bar{\pi}_{it}|R_{it} = 1, \bar{W}_{it}]P(R_{it} = 1|\bar{W}_{it}|X_i\} = E[R_{it}Y_{it}/\bar{\pi}_{it}|X_i]$.

An alternative martingale proof that (2a) implies $E[R_{it}Y_{it}/\bar{\pi}_{it}|X_i] = E[Y_{it}|X_i]$ is provided in the proof of Theorem 1, which follows. We shall prove Theorem 1 under the following regularity conditions. Let $H_i^*(\gamma) = (H_i(\beta)^T, S_{\alpha,i}(\alpha)^T)^T$ with $\gamma^T = (\beta^T, \alpha^T)$. Then we assume:

(R.1) β_0 and α_0 lie in the interior of compact sets $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.
(R.2) $(\bar{R}_{i(T+1)}^T, \bar{W}_{i(T+1)}^T)^T$, $i = 1, \ldots, n$, are independent and identically distributed.
(R.3) $\bar{\lambda}_{it}(\alpha) > c > 0$ for all $\alpha \in \boldsymbol{\alpha}$, $t = (1, \ldots, T)$ for some c.
(R.4) $E[H_i^*(\gamma)] \neq 0$ if $\gamma \neq \gamma_0$.
(R.5) $\text{var}[H_i^*(\gamma_0)]$ is finite and positive definite.
(R.6) $E[\partial H_i^*(\gamma_0)/\partial \gamma^T]$ exists and is invertible.
(R.7) $E[\sup_{\gamma \in \boldsymbol{\gamma}} \|H_i^*(\gamma)\|]$, $E[\sup_{\gamma \in \boldsymbol{\gamma}} \|\partial H_i^*(\gamma)/\partial \gamma^T\|]$, $E[\sup_{\gamma \in \boldsymbol{\gamma}} \|H_i^*(\gamma)H_i^*(\gamma)^T\|]$ are all finite where $\|A\| \equiv \{\Sigma_{ij} A_{ij}^2\}^{1/2}$ for any matrix A with elements A_{ij}, and $\boldsymbol{\gamma}$ is the Cartesian product of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.
(R.8) $L_i(\gamma)$ is a regular parametric model for the observed data, where $L_i(\gamma)$ is a density that differs from the true density only in that γ replaces γ_0.
(R.9) For all γ^* in a neighborhood N of γ_0, $E_{\gamma^*}[H_i^*(\gamma)]$ and $E_{\gamma^*}[\sup_{\gamma \in N} \|H_i^*(\gamma)^T H_i^*(\gamma)\|]$ are bounded, where E_{γ^*} refers to expectation with respect to $L_i(\gamma^*)$.

Throughout, unless unclear we write $U_i = U_i(\beta_0, \alpha_0)$, $U_i^\dagger(\phi) = U_i^\dagger(\beta_0, \phi, \alpha_0)$, $S_{\alpha,i} = S_{\alpha,i}(\alpha_0)$. Similarly, $\partial U_i/\partial\beta$, $\partial U_i^\dagger(\phi)/\partial\beta$, $\partial U_i/\partial\alpha$, $\partial U_i^\dagger(\phi)/\partial\alpha$, and $\partial S_{\alpha,i}/\partial\alpha$ denote evaluation of the derivatives at (β_0, α_0).

Proof of Theorem 1

We first prove that $E(U_i) = 0$. Consider the identity

$$\Delta_i(\alpha) = Id + \sum_{t=1}^{T} (R_{it} - \bar{\lambda}_{it}(\alpha)R_{i(t-1)})\bar{\pi}_{it}(\alpha)^{-1} I^{(t)}, \quad \text{(A.1)}$$

where Id is the $T \times T$ identity matrix and $I^{(t)} = \text{diag}(e_j^{(t)})$ is the $T \times T$ diagonal matrix with diagonal elements $e_j^{(t)}$ satisfying $e_j^{(t)} = 1$ if $j \geq t$ and $e_j^{(t)} = 0$ if $j < t$. Premultiplying by $D_i(\beta)$ and postmultiplying by $\varepsilon_i(\beta)$, we get

$$U_i(\beta, \alpha) = H_i(\beta) + \sum_{t=1}^{T} (R_{it} - \bar{\lambda}_{it}(\alpha)R_{i(t-1)})\bar{\pi}_{it}(\alpha)^{-1} H_{it}(\beta),$$

where $H_{it}(\beta) = D_i(\beta)I^{(t)}\varepsilon_i(\beta)$. Thus $U_i(\beta, \alpha) = H_i(\beta) + M_i(T, \beta, \alpha)$, where

$$M_i(t, \beta, \alpha) = \sum_{j=1}^{t} (R_{ij} - \bar{\lambda}_{ij}(\alpha)R_{i(j-1)})\bar{\pi}_{ij}(\alpha)^{-1} H_{ij}(\beta).$$

We now show that $M_i(t, \beta, \alpha_0)$ is a discrete mean zero martingale with respect to the filtration process $F_i(t) = \sigma\{\bar{W}_{it}, R_{i1}, \ldots, R_{i(t-1)}, Y_i\}$. Note that $Y_i = (Y_{i0}, Y_{i1}, \ldots, Y_{iT})^T$ in $F_i(t)$ includes Y's occurring at times exceeding t. The random variables $N_i(t) = 1 - R_{it}$, for $t = 1, \ldots, T$ define a subject-specific discrete counting process for nonresponse. In view of (2a), $E\{N_i(t)|\bar{W}_{it}, R_{i1}, \ldots, R_{i(t-1)}, Y_i\} = 1 - \bar{\lambda}_{it}R_{i(t-1)}$, thus

$$Z_i(t) = \sum_{j=1}^{t} (R_{ij} - \bar{\lambda}_{ij}R_{i(j-1)})$$

is a mean zero discrete-time martingale process with respect to $F_i(t)$. But

Downloaded by [Virginia Tech Libraries] at 04:07 17 July 2013

$$M_i(t, \beta, \alpha_0) = \sum_{j=1}^{t} \Delta Z_i(j)\bar{\pi}_{ij}(\alpha_0)^{-1}H_{ij}(\beta),$$

where $\Delta Z_i(j) = Z_i(j) - Z_i(j-1)$, so it follows that $M_i(t, \beta, \alpha_0)$ is also a mean zero martingale with respect to $F_i(t)$, because $\bar{\pi}_{it}(\alpha_0)^{-1}H_{it}(\beta)$ is $F_i(t)$ predictable. Similarly, $\partial M_i(t, \beta, \alpha_0)/\partial\beta$ is also a mean zero discrete-time martingale. Thus because, by (1), $E[H_i(\beta_0)] = 0$, we conclude that $E(U_i) = 0$.

As argued later, our regularity conditions guarantee the validity of the standard Taylor expansions

$$n^{1/2}(\hat{\alpha} - \alpha_0) = -E(\partial S_{\alpha i}/\partial\alpha^T)^{-1}S_\alpha(\alpha_0) + o_p(1) \quad \text{(A.2)}$$

and

$$n^{1/2}(\hat{\beta} - \beta_0) = -E(\partial U_i/\partial\beta)^{-1}\{U(\beta_0, \alpha_0) + E(\partial U_i/\partial\alpha^T)^{-1}n^{1/2}(\hat{\alpha} - \alpha_0)\} + o_p(1), \quad \text{(A.3)}$$

where $S_\alpha(\alpha) = n^{-1/2}\Sigma S_{\alpha,i}(\alpha)$ and $U(\beta, \alpha) = n^{-1/2}\Sigma U_i(\beta, \alpha)$. To obtain the limiting distribution of $n^{1/2}\{\hat{\beta} - \beta_0\}$, note that by differentiating $E_{\beta_0,\alpha}\{U_i(\beta_0, \alpha)\} = 0$ with respect to α under the integral sign and evaluating at α_0, we obtain the "generalized information equality" that $E(\partial U_i/\partial\alpha^T) = -E(U_iS_{\alpha i}^T)$ (Pierce 1982). Similarly, $E(\partial S_{\alpha i}/\partial\alpha^T) = -\text{var}(S_{\alpha i})$. Substituting these identities into (A.2) and (A.3), and then substituting the right side of (A.2) for $n^{1/2}(\hat{\alpha} - \alpha_0)$ in (A.3), we obtain

$$n^{1/2}\{\hat{\beta} - \beta_0\} = -\Gamma^{-1}n^{-1/2}\Sigma_i\text{Resid}(U_i, S_{\alpha i}) + o_p(1), \quad \text{(A.4)}$$

and the asymptotic distribution of $n^{1/2}\{\hat{\beta} - \beta_0\}$ follows from (A.4) and the central limit theorem with $\text{var}\{\text{Resid}(U_i, S_{\alpha,i})\} = I - \text{B}\Omega\text{B}^T$. The consistency of $\hat{\Gamma}$, $\hat{I}$, and $\hat{B}$ is a consequence of the law of large numbers. It remains only to justify (A.2) and (A.3). Let $H_i^{**}(\gamma) = (U_i(\beta, \alpha)^T, S_{\alpha,i}(\alpha)^T)^T$, $\gamma^T = (\beta^T, \alpha^T)$. Then, by $\bar{\lambda}_{it}(\alpha)$ bounded below by c, regularity conditions (R.1)–(R.7) remain true for $H_i^{**}(\gamma)$. Further, by $E[U_i(\beta_0, \alpha_0)] = E[D_i(\beta_0)\epsilon_i(\beta_0)] = 0$, $E[H_i^{**}(\gamma_0)] = 0$. But $\hat{\gamma}$ solving $\Sigma_iH_i^{**}(\gamma) = 0$ is precisely $(\hat{\beta}, \hat{\alpha})$. Theorem (3.4) of Newey and McFadden (1993) implies that if $H_i^{**}(\gamma)$ satisfies (R.1)–(R.7) and $E[H_i^{**}(\gamma_0)] = 0$, then, with probability approaching 1, $\hat{\gamma}$ exists and is unique and satisfies

$$n^{1/2}(\hat{\gamma} - \gamma_0) = -E[\partial H_i^{**}(\gamma_0)/\partial\gamma^T]\bar{n}^{1/2} \times \Sigma_iH_i^{**}(\gamma_0) + o_p(1). \quad \text{(A.5)}$$

But (A.5) is precisely Equations (A.2) and (A.3). Regularity conditions (R.8) and (R.9) are needed to justify differentiation under the integral sign.

Proof of Theorem 2

We first derive the properties of $\tilde{\beta}^\dagger$. Redefine $H_i^{**}(\gamma)$ to be $(U_i^\dagger(\beta, \phi, \alpha)^T, S_{\alpha,i}(\alpha)^T)^T$, with $S_{\alpha,i}(\alpha)$ as redefined in Section 6. Then, by $\bar{\lambda}_{it}(\alpha) > c$, regularity conditions (R.1)–(R.7) hold for $H_i^{**}(\gamma)$. Also, $E[H_i^{**}(\gamma_0)] = 0$ because $E[U_i(\beta_0, \alpha_0)] = 0$ and $E[A_i(\phi, \alpha_0)] = 0$ by $E[R_{iT}\{\bar{\pi}_{iT}(\alpha_0)\}^{-1}\Sigma_{r\neq 1}\bar{\pi}_i(r, \alpha_0)\phi_{i,r}|\bar{W}_{i(T+1)}] = \Sigma_{r\neq 1}\bar{\pi}_i(r, \alpha_0)\phi_{i,r} = E[\Sigma_{r\neq 1}I(R_i^* = r)\phi_{i,r}|\bar{W}_{i(T+1)}]$. Thus, by theorem (3.4) of Newey and McFadden (1993), $\hat{\gamma} = (\tilde{\beta}^\dagger, \hat{\alpha})$ solving $\Sigma_iH_i^{**}(\gamma) = 0$ exists and is unique with probability approaching 1 and satisfies (A.5). But (A.5) is equivalent to (A.2) and

$$n^{1/2}(\tilde{\beta}^\dagger - \beta_0) = \{-E[\partial U_i^\dagger(\phi)/\partial\beta^T]\}^{-1} \times \{n^{-1/2}\Sigma_iU_i^\dagger(\phi) - E[\partial U_i^\dagger(\phi)/\partial\alpha^T]n^{1/2}(\hat{\alpha} - \alpha_0)\} + o_p(1). \quad \text{(A.6)}$$

On substituting the right hand side of (A.2) for $n^{1/2}(\hat{\alpha} - \alpha_0)$ in (A.6) and using the generalized information equalities $E[\partial U_i^\dagger(\phi)/\partial\alpha^T] = -E[U_i^\dagger(\phi)S_{\alpha,i}^T] = -E[U_iS_{\alpha,i}^T] + \theta E[A_i(\phi, \alpha_0)S_{\alpha,i}^T]$ and $E[\partial S_{\alpha,i}/\partial\alpha^T] = -\text{var}(S_{\alpha,i})$, we obtain (15). Finally, $n^{1/2}(\tilde{\beta} - \beta_0)$ admits the same expression (15) as $n^{1/2}(\tilde{\beta}^\dagger - \beta_0)$, because with $A(\phi, \hat{\alpha}) = n^{-1/2}\Sigma_iA_i(\phi, \hat{\alpha})$,

$$U(\beta, \hat{\alpha}) - \hat{\theta}A(\phi, \hat{\alpha}) = U(\beta, \hat{\alpha}) - \theta A(\phi, \hat{\alpha}) + o_p(1). \quad \text{(A.7)}$$

Equation (A.7) is true because $n^{1/2}(\hat{\theta} - \theta) = O_p(1)$ and $n^{-1/2}A(\phi, \hat{\alpha}) = o_p(1)$ by $E[A_i(\phi, \alpha_0)] = 0$, implying that $(\hat{\theta} - \theta)A(\phi, \hat{\alpha}) = o_p(1)$.

Proof of Lemma 1

$\text{var}(U_i - S_i)$ minimizes $\text{var}(U_i - G_i)$ over all $G_i \in \mathbf{G} = \{\sum_{t=1}^{T}(R_{it} - \bar{\lambda}_{it}R_{i(t-1)})g_t^*(\bar{W}_{it}), g_t^*(\cdot)$ an arbitrary $p \times 1$ vector-valued function$\}$, because (a) $\text{var}(U_i - \mathbf{G}_i) = \text{var}(U_i - S_i) + \text{var}(S_i - G_i) + E\{(U_i - S_i)(S_i - G_i)^T\} + [E\{(U_i - S_i)(S_i - G_i)^T\}]^T$; (b) S_i and G_i are $F_i(t)$ – adapted martingales; and (c) $E[(U_i - S_i)(S_i - G_i)^T] = 0$ by a martingale covariance calculation for all $G_i \in \mathbf{G}$. Thus, because $U_i - \text{Resid}(U_i, S_{\alpha,i}) \in \mathbf{G}$, $C = \text{var}\{\text{Resid}(U_i, S_{\alpha,i})\} \geq \text{var}(U_i - S_i)$. Further, by standard least squares theory, $\text{Resid}(U_i, S_{\alpha,i}) = U_i - S_i \Leftrightarrow bS_{\alpha,i} = S_i \Leftrightarrow b\partial$ logit $\bar{\lambda}_{it}(\alpha_0)/\partial\alpha = \bar{Q}_{it}$.

APPENDIX B: PROPERTIES OF THE SWEEP ESTIMATOR UNDER NONNORMALITY

We provide a general definition of the sweep estimator that we then specialize to the "usual" and "extended" sweep estimators discussed in the text. Throughout, we assume monotone nonresponse. Let $Z_{it} = (Z_{it1}, \ldots, Z_{its})^T$ be a s-dimensional random vector, $t = 0, \ldots, T$, and let I_s be the $s \times s$ identity matrix. It will be convenient to define $\bar{Z}_{it}^\dagger = (X_i^T, Z_{i0}^T, \ldots, Z_{i(t-1)}^T)^T$ and $\bar{Z}_{i0}^\dagger = X_i$. Note $\dim(\bar{Z}_{it}^\dagger) = p + st$, where p is the dimension of X_i. Let $\hat{b}_t = (\hat{b}_{t1}, \ldots, \hat{b}_{ts})^T$ be the $s \times (p + st)$-dimensional least squares estimate from the multivariate regression of Z_{it} on $\bar{Z}_{it}^\dagger$ among subjects with $R_{it} = 1$. Let $\hat{B}_t$ denote the $[p + s(t+1)] \times (p + st)$-dimensional matrix $(I_{p+st}, \hat{b}_t^T)^T$. The sweep estimator (Little and Rubin 1987) of $E(\bar{Z}_{i(T+1)}^\dagger|X_i)$ is equal to $\hat{\beta}_{\text{sweep},T+1}X_i$, where $\hat{\beta}_{\text{sweep},t} = \hat{B}_{t-1}\hat{B}_{t-2}\ldots\hat{B}_0$ is a $(st + p) \times p$ matrix. When the data are missing at random, $\hat{\beta}_{\text{sweep},T+1}$ is the MLE of $\beta_{0,T+1}$ in the homoscedastic model that assumes, given X_i, $\bar{Z}_{i(T+1)}^\dagger$ is normally distributed with mean $\beta_{0,T+1}X_i$ and unknown covariance matrix Ω_0 (Little and Rubin 1987). (Note that by our notational convention, the first p components of $\bar{Z}_{i(T+1)}^\dagger$ are X_i and thus will have a degenerate distribution given X_i.)

The usual sweep estimator of row 3 of Table 2 uses $Z_{it} = Y_{it}$. The extended sweep estimator of row 9 of Table 2 uses $Z_{it} = (Y_{it}, V_{it})$, and that of row 10 uses $Z_{it} = (Y_{it}, V_{it}^{.33})$. All use $X_i = 1$. The following theorem provides sufficient conditions for the general sweep estimator to be asymptotically normal and unbiased.

Theorem B.1. If for $t = 1, \ldots, T$,

$$E(Z_{it}|\bar{Z}_{it}^\dagger, R_{i(t-1)} = 1) = b_t\bar{Z}_{it}^\dagger \quad \text{(A.8)}$$

for some $s \times (p + st)$ matrix b_t,

$$P(R_{it} = 1|R_{i(t-1)} = 1, \bar{Z}_{i(T+1)}^\dagger) = P(R_{it} = 1|R_{i(t-1)} = 1, \bar{Z}_{it}^\dagger), \quad \text{(A.9)}$$

Downloaded by [Virginia Tech Libraries] at 04:07 17 July 2013

and

$$P(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{Z}^{\dagger}_{it}) > \sigma > 0, \qquad (A.10)$$

then (a)

$$E(Z_{it} \mid \bar{Z}^{\dagger}_{it}, R_{it} = 1) = E(Z_{it} \mid \bar{Z}^{\dagger}_{it}), \qquad (A.11)$$

(b) $E(\bar{Z}^{\dagger}_{i(T+1)} \mid X_i) = \beta_{0,T+1} X_i$, where $\beta_{0,t} = B_{t-1} \ldots B_1 B_0$ and B_t is defined like $\hat{B}_t$ but with b_t replacing $\hat{b}_t$; (c) $n^{1/2}(\hat{\beta}_{\text{sweep},T+1} - \beta_{0,T+1})$ is asymptotically normal with mean zero; and (d) when $X_i = 1$, and thus $p = 1$, a consistent variance estimator is $\hat{V}_{\text{sweep},T+1}$, recursively defined by $\hat{V}_{\text{sweep},t} = n\hat{\nu}_{t-1}/\Sigma_i R_{i(t-1)} + \hat{B}_{t-1}\hat{V}_{\text{sweep},t-1}\hat{B}^T_{t-1}$ with $\hat{V}_{\text{sweep},0} = 0$ and $\hat{\nu}_t$ being the $p + s(t+1)$ square matrix all of whose entries are zero except for the $s \times s$ lower right submatrix $\hat{\mu}_t$, where $\hat{\mu}_t$ has r, r' entry $\hat{\beta}^T_{\text{sweep},t}\hat{\sigma}_{t,rr'}\hat{\beta}_{\text{sweep},t}$ with $\hat{\sigma}_{t,rr'} = (\hat{\sigma}^*_t)^{-1}\hat{\sigma}^{\dagger}_{t,rr'}(\hat{\sigma}^*_t)^{-1}$, with $\hat{\sigma}^*_t = \{\Sigma_i R_{it}\}^{-1}\Sigma_i R_{it}\bar{Z}^{\dagger\otimes 2}_{it}$, $\hat{\sigma}^{\dagger}_{t,rr'} = \{\Sigma_i R_{it}\}^{-1}\Sigma_i R_{it}\hat{\rho}_{itr}\hat{\rho}^T_{itr'}$ with $\hat{\rho}_{itr} = \bar{Z}^{\dagger}_{it}\{Z_{itr} - \hat{b}_{tr}\bar{Z}^{\dagger}_{it}\}$. [$\hat{\sigma}_{t,rr'}$ is a robust estimator of the asymptotic covariance of $(\Sigma_i R_{it})^{1/2}(\hat{b}_{tr} - b_{tr})$ and $(\Sigma_i R_{it})^{1/2}(\hat{b}_{tr'} - b_{tr'})$]. A similar recursive formula for the asymptotic variance of $n^{1/2}(\hat{\beta}_{\text{sweep},T+1} - \beta_{0,T+1})$ can be obtained when X_i is multivariate by using "vec" notation.

With $Z_{it} = (Y_{it}, V^T_{it})^T$, according to Theorem B.1, Equation (A.8) implies that the sweep estimators of both $E(\bar{Y}_{i(T+1)} \mid X_i)$ and of $E(\bar{V}_{i(T+1)} \mid X_i)$ are consistent when Equation (2b) and the missing at random assumption Equation (3) hold. Note that when Equation (3) is false, $E(\bar{V}_{i(T+1)} \mid X_i)$ is not identified under the weaker assumption (2a); however, the following theorem gives sufficient conditions for the consistency of the sweep estimator of $E(\bar{Y}_{i(T+1)} \mid X_i)$.

Theorem B.2. Suppose that Y_{it} is a subvector of Z_{it}. If (A.8), and (A.10) are true, and for all $t' \geq t$,

$$E(Y_{it'} \mid R_{it} = 1, \bar{Z}_{it}) = E(Y_{it'} \mid R_{i(t-1)} = 1, \bar{Z}_{it}), \qquad (A.12)$$

then (a) $E(\bar{Y}_{i(T+1)} \mid X_i) = \beta^Y_{0,T+1} X_i$, where $\beta^Y_{0,T+1}$ is the submatrix of $\beta_{0,T+1}$ with rows corresponding to the sub-vector $\bar{Y}_{i(T+1)}$ of $\bar{Z}_{i(T+1)}$; and (b) $\hat{\beta}^Y_{\text{sweep},T+1}$, the corresponding subvector of $\hat{\beta}_{\text{sweep},T+1}$, is asymptotically normal and unbiased for $\beta^Y_{0,T+1}$ with asymptotic variance that, when $X_i = 1$, can be consistently estimated by the corresponding submatrix of $\hat{V}_{\text{sweep},T+1}$.

Note that Equation (2a) implies (A.12). Also note that $\hat{\beta}^Y_{\text{sweep},T+1} X_i$ is consistent for $E(\bar{Y}_{i(T+1)} \mid X_i)$ even if (A.11) is false, provided that (A.8), (A.10), and (A.12) are true. The key step in the proof of Theorems B.1 and B.2 is the following lemma. Set $\eta_{i(t'+1)t'} = Y_{it'}$, and recursively define $\eta_{itt'} = E(\eta_{i(t+1)t'} \mid \bar{Z}^{\dagger}_{it}, R_{it} = 1)$ for $t = t', \ldots, 0$.

Lemma B.1. If (A.10) and (A.12) are true, then $E[Y_{it'} \mid \bar{Z}^{\dagger}_{it}, R_{it} = 1] = \eta_{itt'}$, $t = (0, \ldots, t')$. In particular, $E[Y_{it'} \mid X_i] = \eta_{i0t'}$. The proof of Lemma B.1 is a straightforward exercise in repeated iterated expectations. It also follows from the G-computation algorithm formula (5a).

Proof of Theorems B.2 and B.1

Consider first Theorem B.2. By recursively evaluating $\eta_{it't'}$, $\eta_{i(t'-1)t'}, \ldots, \eta_{i0t'}$ for $t' = (0, \ldots, T)$ under (A.8), part (a) of Theorem B.2 follows immediately from Lemma B.1. To prove part (b), note by standard least squares theory and Equation (A.8), $(\Sigma_i R_{it})^{1/2}(\hat{b}_t - b_t)$ is asymptotically normal with a consistent variance estimator that has r, r' entry $\hat{\sigma}_{t,rr'}$. Thus $n^{1/2}(\hat{\beta}^Y_{\text{sweep},T+1} - \beta^Y_{0,T+1})$ is asymptotically normal and unbiased by Slutsky's theorem. The formula for the estimator of $\text{var}^A\{n^{1/2}(\hat{\beta}^Y_{\text{sweep},T+1} - \beta^Y_{0,T+1})\}$ is a recursive calculation using the fact that $\hat{B}_{t+1}$ is random but $\hat{B}_t\hat{B}_{t-1} \ldots \hat{B}_0$ are fixed given $\{\bar{Z}^{\dagger}_{i(t+1)}, \bar{R}_{i(t+1)}; i = 1, \ldots, n\}$. Turn now to Theorem B.1. Parts (b) and (c) of Theorem B.1 are special cases of Theorem B.2 with $Z_{it} = Y_{it}$. To prove part (a), one can calculate that (A.9) and (A.12) imply (A.11).

A "Nonlinear" Sweep Estimator

Even when the assumption (A.8) of linear conditional expectations fails, we can use Lemma B.1 to motivate the following "nonlinear" sweep (or iterated conditional expectations) estimators of $E[Y_{it'} \mid X_i]$. Specify models $\eta_{itt'}(\theta_{tt'}) \equiv \eta_{tt'}(\bar{Z}_{it}, \theta_{tt'})$, $t = (0, \ldots, t')$, for the unknown $\eta_{itt'}$, where each $\theta_{tt'}$ is an unknown parameter vector and the $\eta_{itt'}(\cdot, \cdot)$ are known, possibly nonlinear, functions. Estimate $E[Y_{it'} \mid X_i]$ by $\hat{\eta}_{i0t'}$, where $\hat{\eta}_{itt'}$ is recursively defined by (a) $\hat{\eta}_{i(t+1)t'} = Y_{it'}$, and (b) $\hat{\eta}_{itt'}$ equals the predicted values for each subject i with $R_{i(t-1)} = 1$ from the nonlinear least squares regression of $\hat{\eta}_{j(t+1)t'}$ on $\bar{Z}_{jt}$ among subjects j with $R_{jt} = 1$ based on the regression function $\eta_{tt'}(\theta_{tt'})$. Then, by Lemma B.1 and standard least squares theory, if the models are correctly specified in the sense that for some $\theta_{0t'} = (\theta_{00t'}, \ldots, \theta_{0tt'})$, $\eta_{itt'}(\theta_{0tt'}) = \eta_{itt'}$, $t = (0, \ldots, t')$, then $\hat{\eta}_{i0t'}$ will be asymptotically normal and unbiased for $E[Y_{it'} \mid X_i]$. Further, if we choose the regression function $\eta_{0t'}(X_i, \theta_{0t'})$ to equal the regression function $g_{t'}(X_i, \beta)$ of Equation (1), then our model for $E[Y_{it'} \mid X_i]$ will be consistent with model (1). Thus the "nonlinear" sweep estimator would appear to be a reasonable alternative to a weighted estimator for estimating a single conditional mean $E[Y_{it'} \mid X_i]$. But the nonlinear sweep estimator of $E[Y_{it'} \mid X_i]$, with X_i a dichotomous treatment arm indicator, can be inconsistent even in the absence of missing data or censoring if any of the models $\eta_{itt'}(\theta_{tt'})$, $t = 1, \ldots, t'$ are misspecified, thus violating the spirit of the intent-to-treat principle. Further difficulties arise when trying to simultaneously estimate the conditional mean of T outcomes $Y_{i1}, \ldots, Y_{iT}$ given X_i based on the nonlinear sweep estimators. Specifically, the estimated regression functions $\eta_{tt'}(\bar{Z}^{\dagger}_{it}, \hat{\theta}_{tt'})$ and $\eta_{tt''}(\bar{Z}^{\dagger}_{it}, \hat{\theta}_{tt''})$, where $t' \neq t''$ and $t < \min(t', t'')$, may be incompatible in the sense there exists no distribution for the data that satisfies the joint restriction implied by both regression functions.

[*Received February 1992. Revised March 1994.*]

REFERENCES

Begun, J. M., Hall, W. J., Huang, W. M., and Wellner, J. A. (1983), "Information and Asymptotic Efficiency in Parametric-nonparametric Models," *The Annals of Statistics*, 11, 432–452.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Diggle, P., and Kenward, M. G. (1994), "Informative Dropout in Longitudinal Data Analysis" (with discussion), *Applied Statistics*, 43, 49–94.

Finkelstein, D., and Schoenfeld, D. A. (1994), "Analyzing Survival in the Presence of an Auxiliary Variable," *Statistics in Medicine*, 13, 1747–1754.

Fischl, M. A., Parker, C. B., Pettinelli, C., Wulfsohn, M., Hirsch, M. S., Collier, A. C., Antoniskis, D., Ho, M., Richman, D. D., Fuchs, E., Merigan, T. C., Reichman, R. C., Gold, J., Steigbibel, N., Leoung, G. S., Rasheed, S., Tsiatis, A., and the AIDS Clinical Trials Group (1990), "Randomized Controlled Trial of a Reduced Daily Dose of Zidovudine in Patients With the Acquired Immunodeficiency Syndrome," *New England Journal of Medicine*, 323, 1009–1019.

Flanders, W. D., and Greenland, S. (1991), "Analytic Methods for Two-Stage Case-Control Studies and Other Stratified Designs," *Statistics in Medicine*, 10, 739–747.

Fleming, T. R., Prentice, R. L., Pepe, M. S., and Glidden, D. (1994), "Surrogate and Auxiliary End Points in Clinical Trials, With Potential Applications in Cancer and AIDS Research," *Statistics in Medicine*, 13, 955–968.

Gourieroux, C., Monfort, A., and Trognon, A. (1984), "Pseudo-Maximum Likelihood Methods: Theory," *Econometrica*, 52, 681–700.

Gray, R. (1992), "A Kernel Method for Incorporating Information on Disease Progression in the Analysis of Survival," unpublished manuscript.

Heyting, A., Tolboom, J. T. B. M., and Essers, J. G. A. (1992), "Statistical Handling of Drop-Outs in Longitudinal Clinical Trials," *Statistics in Medicine*, 11, 2043–2062.

Downloaded by [Virginia Tech Libraries] at 04:07 17 July 2013

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association,* 47, 663–685.

Huber, P. (1985), "Projection Pursuit," *The Annals of Statistics,* 13, 435–474.

Kalbfleisch, J. D., and Lawless, J. F. (1988), "Likelihood Analysis of Multi-Stage Models for Disease Incidence and Mortality," *Statistics in Medicine,* 7, 149–160.

Kalbfleisch, J. D., and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data,* New York: John Wiley.

Lagakos, S. W. (1977), "Using Auxiliary Variables for Improved Estimates of Survival Time," *Biometrics,* 33, 399–404.

Landwehr, J. M., Pregibon, D., and Shoemaker, A. C. (1984), "Graphical Methods for Assessing Logistic Regression Models," *Journal of the American Statistical Association,* 79, 61–71.

Liang, K.-Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika,* 73, 13–22.

Lin, D. Y., Robins, J. M., Wei, L. J. (1995), "Comparing Two Failure Time Distributions in the Presence of Informative Censoring," submitted to *Biometrika.*

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

Malani, H. M. (1992), "Modification of the Redistribution to the Right Algorithm Using Disease Markers, Technical Report 3, Biostatistics Technical Reports, University of California, Berkeley.

Manski, C. F. (1988), *Analog Estimation Methods in Econometrics,* New York: Chapman and Hall.

Manski, C. F., and Lerman, S. (1977), "The Estimation of Choice Probabilities From Choice-Based Samples," *Econometrica,* 45, 1977–1988.

Mark, S. D., and Gail, M. H. (1994), "A Comparison of Likelihood-Based Marginal Estimating Equation Methods for Analyzing Repeated-Ordered Categorical Responses With Missing Data: Application to an Intervention Trial of Vitamin Prophylaxis for Esophageal Dysplasia," *Statistics in Medicine,* 13, 479–494.

Mark, S. D., and Robins, J. M. (1993a), "Estimating the Causal Effect of Smoking Cessation in the Presence of Confounding Factors Using a Rank-Preserving Structural Failure Time Model," *Statistics in Medicine,* 12, 1605–1628.

——— (1993b), "A Method for the Analysis of Randomized Trials With Compliance Information: An Application to the Multiple Risk Factor Intervention Trial," *Controlled Clinical Trials,* 14, 79–97.

McCullagh, R., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Newey, W. K. (1994), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica* (forthcoming).

Newey, W. K., and McFadden, D. (1993), "Estimation in Large Samples," *Handbook of Econometrics,* Vol. 4, eds. D. McFadden and R. Engler, Amsterdam: North-Holland (forthcoming).

Pearl, J. (1994), "Causal Diagrams for Empirical Research," submitted to *Biometrika.*

Pierce, D. A. (1982), "The Asymptotic Effect of Substituting Estimators for Parameters in Certain Types of Statistics," *The Annals of Statistics,* 10, 475–478.

Prentice, R. L. (1988), "Correlated Binary Regression With Covariates Specific to Each Binary Observation," *Biometrics,* 44, 1033–1048.

Robins, J. M. (1986), "A New Approach to Causal Inference in Mortality Studies With Sustained Exposure Periods—Application to Control of the Healthy Worker Survivor Effect," *Math Modeling,* 7, 1393–1512.

——— (1987a), Addendum to "A New Approach to Causal Inference in Mortality Studies With Sustained Exposure Periods—Applications to Control of the Healthy Worker Survivor Effect," *Computers and Mathematics With Application,* 14, 923–945.

——— (1987b), "A Graphical Approach to the Identification and Estimation of Causal Parameters in Mortality Studies With Sustained Exposure Periods," *Journal of Chronic Disease,* 40, Suppl. 2, 139s–161s.

——— (1989), "The Analysis of Randomized and Nonrandomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," in *Health Service Research Methodology: A Focus on AIDS,* eds. L. Sechrest, H. Freeman and A. Mulley, NCHSR, U.S. Public Health Service, pp. 113–159.

——— (1991), "Estimating Regression Coefficients in the Presence of Dependent Censoring" submitted to *Journal of the American Statistical Association.*

——— (1993a), "Analytic Methods for HIV Treatment and Cofactor Effects," in: *Methodological Issues of AIDS Behavioral Research,* eds. D. G. Ostrow and R. Kessler, New York: Plenum Press, pp. 213–287.

——— (1993b), "Information Recovery and Bias Adjustment in Proportional Hazards Regression Analysis of Randomized Trials Using Surrogate Markers," in *Proceedings of the Biopharmaceutical Section, American Statistical Association,* pp. 24–33.

——— (1994a), "Correcting for Noncompliance in Randomized Trials Using Structural Nested Mean Models," *Communications in Statistics,* pp. 2379–2412.

——— (1994b), "An Analytic Method for Randomized Trials with Informative Censoring," submitted to *Lifetime Data Analysis.*

——— (1995a), "Causal Inference From Complex Longitudinal Data Using Structural Nested Models," in *Latent Variable Modelling with Applications to Causality: Lecture Notes in Statistics,* NY: Springer-Verlag (to appear).

——— (1995b), "Locally Efficient Median Regression with Random Censoring and Surrogate Markers," *Proceedings of the 1994 Conference on Lifetime Data Models in Reliability and Survival Analysis,* Boston, MA. (to appear).

Robins, J. M., and Greenland, S. (1994), "Adjusting for Different Rates of PCP Prophylaxis in High-Dose Versus Low-Dose AZT Treatment Arms in an AIDS Randomized Trial," *Journal of the American Statistical Association,* 89, 427, pp. 737–749.

Robins, J. M., and Rotnitzky, A. (1992), "Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers," in *AIDS Epidemiology—Methodological Issues,* eds. N. Jewell, K. Dietz, and V. Farewell, Boston: Birkhäuser, pp. 297–331.

——— (1995), "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association,* 90, 122–129.

——— (1995), "Estimation in Semi- and Non-parametric Models with Non-ignorable Missingness," *Technical Report,* Harvard School of Public Health, Department of Epidemiology.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors are not Always Observed," *Journal of the American Statistical Association,* 89, 427, pp. 846–866.

Rosenbaum, P. R. (1987), "Model-Based Direct Adjustment," *Journal of the American Statistical Association,* 82, 387–394.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika,* 63, 581–92.

Spirtes, P., Glymour, C., and Scheines, R. (1993), *Causation, Prediction, and Search,* New York: Springer Verlag.

Zhao, L. P., and Lipsitz, S. (1992), "Design and Analysis of Two-Stage Studies," *Statistics in Medicine,* 11, 769–782.

Downloaded by [Virginia Tech Libraries] at 04:07 17 July 2013