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

Semiparametric Efficiency in Multivariate Regression Models with Missing Data

James M. Robins [a] & Andrea Rotnitzky [b]
[a] Epidemiology and Biostatistics, Harvard University, Boston, MA, 02115
[b] Biostatistics, Harvard University, Boston, MA, 02115
Published online: 27 Feb 2012.

To cite this article: James M. Robins & Andrea Rotnitzky (1995) Semiparametric Efficiency in Multivariate Regression Models with Missing Data, Journal of the American Statistical Association, 90:429, 122-129

To link to this article: http://dx.doi.org/10.1080/01621459.1995.10476494

Semiparametric Efficiency in Multivariate Regression Models with Missing Data

James M. ROBINS and Andrea ROTNITZKY*

We consider the efficiency bound for the estimation of the parameters of semiparametric models defined solely by restrictions on the means of a vector of correlated outcomes, Y, when the data on Y are missing at random. We show that the semiparametric variance bound is the asymptotic variance of the optimal estimator in a class of inverse probability of censoring weighted estimators and that this bound is unchanged if the data are missing completely at random. For this case we study the asymptotic performance of the generalized estimating equations (GEE) estimators of mean parameters and show that the optimal GEE estimator is inefficient except for special cases. The optimal weighted estimator depends on unknown population quantities. But for monotone missing data, we propose an adaptive estimator whose asymptotic variance can achieve the bound.

KEY WORDS: Correlated outcomes; Generalized estimating equations; Generalized least squares; Missing at random; Longitudinal studies.

1. INTRODUCTION

We consider a longitudinal study conducted over a fixed interval from time 1 to T. Let $Y_i = (Y_{i1}, \ldots, Y_{iT})^T$, $i = 1, \ldots, n$, be the vector of multivariate outcomes corresponding to the ith subject measured at prespecified visit times $(1, \ldots, T)$. Here and throughout, T, when used as a superscript, denotes matrix transposition. We suppose that the goal of the study is to describe the evolution of the mean of Y_{it}, $t = 1, \ldots, T$, conditional on a vector of variables X_i. In randomized studies, X_i may include a treatment arm indicator and pretreatment variables such as age, sex, and race.

Extensive literature exists on the efficient estimation of the parameters β_0 of the regression of Y_i on X_i in the absence of missing data. When the mean of Y_i is $X_i^T\beta_0$ and the covariance of Y_i given X_i is known, the generalized least squares estimator $\hat{\beta}_G$ of β_0 is best linear unbiased (Rao 1973, p. 301). With unknown conditional covariance matrix, $\hat{\beta}_G$ is not a feasible estimator, because it depends on the unknown covariance function. But Carroll and Ruppert (1982) and Robinson (1988) showed that if the covariance matrix is a smooth function of X_i, then the generalized least squares estimator $\tilde{\beta}_G$ that replaces the unknown covariance matrix by a nonparametric estimator has the same asymptotic distribution as $\hat{\beta}_G$. Furthermore, Chamberlain (1987) showed that the asymptotic variance of $\hat{\beta}_G$ attains the semiparametric variance bound for regular estimators of β_0 in the semiparametric model defined solely by the linear model restrictions on the marginal means. The efficiency of estimators of β_0 has also been studied in settings where the mean of Y_{it} is linked to X_i through some known arbitrary function $g_t(X_i, \beta)$. Chamberlain (1987) showed that the semiparametric variance bound under a semiparametric model that imposes restrictions solely on the conditional means of Y_{it} given X_i is achieved at the asymptotic variance of the solution to estimating equations that are the multivariate version of the quasi-likelihood equations introduced by Wedderburn (1976) and studied by McCullagh (1983). These estimating equations are in the class of generalized estimating equations (GEE) discussed by Liang and Zeger (1986) and Gourieroux, Monfort, and Trognon (1984). As in the linear case, semiparametric feasible efficient estimators of β_0 can be obtained, under smoothness conditions, by substituting nonparametric estimates of the unknown conditional covariance matrix into the optimal estimating equations (Newey 1991).

The goal of this article is to extend Chamberlain's efficiency results to the setting in which some Y_{it} are missing. We consider several missing data (i.e., nonresponse) mechanisms. The first mechanism assumes that the data are missing completely at random (Rubin 1976); that is, the missing data process may depend on the covariates X_i but is otherwise independent of observed and unobserved outcomes. A second mechanism allows for the probability of nonresponse at visit (i.e., occasion) t to depend on a subject's observed past, including the past history of a vector of time-dependent covariates V_{it} that are correlated with Y_i. The first mechanism is a special case of the second. We derive the efficient score equations; that is, the estimating equations that have a solution whose asymptotic variance coincides with the semiparametric variance bound. We show that the form of the bound is the same under both missing data mechanisms considered.

In Section 2 we present the model and we review a class of estimators proposed by Robins, Rotnitzky, and Zhao (1995), denoted hereon as the RRZ class, that extends the class of weighted least squares estimators to setting where nonresponse depends on the past. The RRZ estimators require a model for the nonresponse probabilities. In Section 3 we calculate the semiparametric variance bound for the model and show that this bound is the same whether the nonresponse probabilities are completely known, completely unknown, or follow a model. We further show that the bound is achieved at the asymptotic variance of the optimal estimator in the RRZ class. In Section 4 we specifically study the case of data missing completely at random. In particular, we study the asymptotic performance of the GEE estimators proposed by Liang and Zeger (1986) and Gourieroux et al.

* James M. Robins is Professor of Epidemiology and Biostatistics and Andrea Rotnitzky is Assistant Professor of Biostatistics, Harvard University, Boston, MA 02115. Support for this research was provided in part by Grants 2 P30 ES00002, R01-AI32475, R01-ES03405, K04-ES00180, GM-48704, and GM-29745 from the National Institutes of Health. Andrea Rotnitzky was additionally supported in part by a Mellon Foundation Faculty Development Award.

Journal of the American Statistical Association
March 1995, Vol. 90, No. 429, Theory and Methods

(1984) and show that the optimal estimator in the GEE class is inefficient except under special conditions. A small simulation study illustrates this theoretical result. Because the solution to the efficient score equations depends on unknown population parameters, in Section 5 we propose a two-stage adaptive RRZ estimator whose asymptotic variance can achieve the semiparametric variance bound and examine its performance in a small simulation study. In Sections 2–5 we assume that the missing data pattern is monotone; that is, once a subject misses a visit, return is not possible. In Section 6 we extend our results to arbitrary missing data patterns. Finally, in Section 7 we present some final remarks.

2. THE MODEL

To define our model, let $Y_i = (Y_{i1}, \ldots, Y_{iT})$ and $X_i = (X_{i0}^T, \ldots, X_{iT}^T)^T$, where X_{it} is a vector of explanatory variables at visit t and visit zero occurs just prior to start of follow-up. Often the X_{it} are deterministic functions of time and the baseline variables X_i^*; for example, $X_{it} = X_i^* t$. Further, suppose 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 $W_{it} = (V_{it}^T, Y_{it})^T$, $t = 1, \ldots, T$, and $W_{i0} = (X_i, V_{i0})$. 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 $R_{i0} = 1$ for all subjects i and that Y_{it} and V_{it} are either both observed or both missing at each time $t \geq 1$. Until Section 5, we shall assume that the missing data patterns are monotone; that is, $R_{it} = 0$ implies $R_{i(t+1)} = 0$. We also shall assume that X_i is completely observed (i.e., known); that is, X_{it} is observed even if $R_{it} = 0$, as would be the case if $X_{it} = X_i^* t$. We assume that $(\bar{W}_{i(T+1)}^T, R_{i1}, \ldots, R_{iT})$, $i = 1, \ldots, n$, are independent and identically distributed, where for any variable z_{it}, $t = 0, \ldots, T$, $\bar{z}_{it} = (z_{i0}^T, \ldots, z_{i(t-1)}^T)^T$, and we use overbars to indicate past data recorded up to but not including the corresponding occasion. Consider the following two conditions on the missing data process:

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

and

$$P(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{i(T+1)}) = P(R_{it} = 1 \mid R_{i(t-1)} = 1, X_i). \quad (2)$$

Under (1), censoring (i.e., nonresponse) is unrelated to current and future outcomes given the past $\bar{W}_{it}$. Assumption (1) holds in particular when the data are missing at random in the sense of Rubin (1976), because under monotonicity, missing at random is equivalent to

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

When assumption (2) holds, we shall say that the data are missing completely at random. Assumption (2) implies both (1) and (3). We suppose that the marginal distribution of Y_{it} given X_i follows the (possibly nonlinear) regression model

$$E(Y_{it} \mid X_i) = g_t(X_i, \beta_0), \quad (4)$$

where β_0 is a $p \times 1$ vector of unknown parameters and $g_t(\cdot, \cdot)$, for $t = 1, \ldots, T$, are known functions. The goal of this article is to study the efficiency with which we can estimate β_0 under nonresponse processes satisfying (1), (2), or (3). It is important to understand that even when data on the time-dependent covariates V_{it} are available, it is the conditional mean of Y_i given X_i alone that it is of substantive interest. To estimate β_0 under dependent censoring when the dimension of $\bar{W}_{it}$ is large and the nonresponse probabilities are unknown, Robins et al. (1995) assumed 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,

$$\bar{\lambda}_{it} = \bar{\lambda}_{it}(\alpha_0), \quad (5)$$

where for each α, $\bar{\lambda}_{it}(\alpha)$ is a known function of $\bar{W}_{it}$ taking values in $(0, 1]$. Given $\hat{\alpha}$, the value of α that maximizes the partial likelihood

$$L(\alpha) = \prod_i \prod_t [\bar{\lambda}_{it}(\alpha)^{R_{it}} \{1 - \bar{\lambda}_{it}(\alpha)\}^{1-R_{it}}]^{R_{i(t-1)}}, \quad (6)$$

they proposed solving the estimating equation

$$\sum_{i=1}^{n} U_i(\beta, \hat{\alpha}) = 0, \quad (7)$$

where $U_i(\beta, \alpha) = D_i(\beta)\Delta_i(\alpha)\varepsilon_i(\beta)$, $\Delta_i(\alpha)$ is the $T \times T$ diagonal matrix with diagonal elements $\Delta_{it}(\alpha) = \bar{\pi}_{it}(\alpha)^{-1} R_{it}$, $\bar{\pi}_{it}(\alpha) = \bar{\lambda}_{i1}(\alpha) \times \cdots \times \bar{\lambda}_{it}(\alpha)$, $\varepsilon_i(\beta) = (\varepsilon_{i1}(\beta), \ldots, \varepsilon_{iT}(\beta))^T$, $\varepsilon_{it}(\beta) = Y_{it} - g_t(X_i, \beta)$, and $D_i(\beta) = d(X_i, \beta)$ is a $p \times T$ matrix of known functions of X_i. Robins et al. (1995) showed that under mild regularity conditions, and provided that (1), (4), (5), and, for $t = 1, \ldots, T$,

$$\bar{\lambda}_{it} > \sigma > 0 \quad (8)$$

hold, the RRZ estimating equations (7) have a solution $\hat{\beta}$ such that $n^{1/2}(\hat{\beta} - \beta_0)$ is asymptotically normal with zero mean. For a fixed choice of $d(\cdot, \cdot)$, they showed in their lemma 1 that the asymptotic variance of $\hat{\beta}$ is greater than or equal to that of the estimator that solves (7) when $\bar{\lambda}_{it}^{(Q)}(\hat{\omega})$ replaces $\bar{\lambda}_{it}(\hat{\alpha})$, where $\hat{\omega}$ is the maximum partial likelihood estimator of ω in the model

$$\text{logit } \bar{\lambda}_{it}^{(Q)}(\omega) = \text{logit } \bar{\lambda}_{it}(\alpha) + \delta^T \bar{Q}_{it}, \quad (9)$$

where $\omega = (\alpha^T, \delta^T)^T$, $\bar{\lambda}_{it}(\cdot)$ satisfies (5) for some α_0; δ_0, the true value of δ, is zero because (5) is true; and $\bar{Q}_{it} = D_i(\beta_0)\bar{Q}_{it}^*$, where $\bar{Q}_{it}^*$ is the T vector with the first $t - 1$ components zero and, for $t \leq j \leq T$, the jth component $\bar{\pi}_{it}^{-1} G_{itj}$ with

$$G_{itj} \equiv E(\varepsilon_{ij} \mid R_{i(t-1)} = 1, \bar{W}_{it}).$$

3. SEMIPARAMETRIC EFFICIENCY

In this section we show that the optimal RRZ estimator attains the semiparametric variance bound. Before stating the main results of this section, we define the semiparametric variance bound, following Begun, Hall, Huang, and Wellner (1983), Bickel, Klaassen, Ritov, and Wellner (1993) and

Newey (1990). Suppose that the data consist of n independent copies Z_i, $i = (1, \ldots, n)$, of a random variable Z. Let $L(\beta, \theta; Z_i)$ be the likelihood for a subject i in a semiparametric model indexed by a $p \times 1$ parameter vector β and a nuisance parameter θ taking values in some infinite-dimensional set. Let (β_0, θ_0) index the distribution generating Z_i. Define a regular parametric submodel to be a regular fully parametric model with parameters (β, η) and likelihood $L(\beta, \eta; Z_i)$ with true values (β_0, η_0) where the "sub" prefix refers to the fact that for each η, the distribution $L(\beta, \eta; Z_i)$ is a distribution $L(\beta, \theta; Z_i)$ allowed by the semiparametric model. An estimator of β_0 is regular in a regular parametric submodel if locally it converges uniformly to its limiting distribution. A regular estimator of β_0 is an estimator that is regular in every regular parametric submodel. The semiparametric variance bound for β_0 is the supremum of the Cramer–Rao variance bounds for β_0 over all regular parametric submodels. The asymptotic variance of any regular estimator of β_0 is no smaller than the semiparametric variance bound. In the context of this article, β_0 is the parameter in Equation (4). Further, for a subject i lost to follow-up at t (i.e., $R_{i(t-1)} = 1$ and $R_{it} = 0$), Z_i equals $(\bar{W}_{it}^T, R_i^T)^T$, where $R_i = (R_{i1}, \ldots, R_{iT})^T$, and for subject i not lost to follow-up, Z_i equals $(\bar{W}_{i(T+1)}^T, R_i^T)$. The following theorem, proved in Appendix A, provides the semiparametric variance bound in the semiparametric model defined by restrictions (3), (4), and (8) and states that the bound is attained at the asymptotic variance of a solution to an estimating equation of the form (7).

Theorem 1. The asymptotic variance of the solution $\hat{\beta}_{op}$ of Equation (7) that uses $D_{op,i}(\beta)$ and $\bar{\lambda}_{it}^{(Q)}(\hat{\omega})$ instead of $D_i(\beta)$ and $\bar{\lambda}_{it}(\hat{\alpha})$ attains the semiparametric variance bound for regular estimators of β_0 in the semiparametric model (a) defined by (3), (4), and (8), where $D_{op,i}(\beta) = \{\partial g(X_i, \beta)/\partial\beta\}\{E[(U_i^* - P_i^*)^{\otimes 2}|X_i]\}^{-1}$, with $U_i^* = \Delta_i\varepsilon_i$, $P_i^* = \sum_{t=1}^T (R_{it} - \bar{\lambda}_{it}R_{i(t-1)})\bar{Q}_{it}^*$, $\varepsilon_i = \varepsilon_i(\beta_0)$, $\Delta_i = \Delta_i(\alpha_0)$, and $A^{\otimes 2} = AA^T$. Furthermore, $\Gamma_{opt}^{-1} = -E\{\partial D_{op,i}(\beta_0)\varepsilon_i(\beta_0)/\partial\beta^T\}^{-1}$ is the semiparametric variance bound.

Theorem 2 states that the semiparametric variance bound is unchanged if in the semiparametric model defined by (3), (4), and (8), restriction (3) is replaced by the weaker condition (1) or the yet-weaker condition $E[Y_{it}|R_{ij} = 1, \bar{W}_{ij}] = E[Y_{it}|R_{i(j-1)} = 1, \bar{W}_{ij}]$, $t \geq j$. It additionally states that Γ_{opt}^{-1} is also the bound in the semiparametric models defined by adding to (1), (4), and (8) restrictions on the nonresponse probabilities $\bar{\lambda}_{it}$. Thus under (4) and (8) and when either (1) or (3) hold, prior knowledge concerning the probabilities $\bar{\lambda}_{it}$ does not provide further information about β_0. In particular, this implies that given (1) or (3), knowledge that the nonresponse process satisfies the stronger condition (2) that the data are missing completely at random does not provide additional information about β_0 because Equation (2) is equivalent to Equation (3) plus a restriction on $\bar{\lambda}_{it}$.

Theorem 2. Γ_{opt}^{-1} is also the semiparametric variance bound in the ten additional models (b)–(k) characterized by the restrictions (4), (8), and (b), (1); (c), (1) and $\bar{\lambda}_{it}$, a known function of the $\bar{W}_{it}$; (d), (1) and (5); (e), (3) and $\bar{\lambda}_{it}$, a known function of the $\bar{W}_{it}$; (f), (3) and (5); (g), (2); (h), (2) and $\bar{\lambda}_{it}$, a known function of the $\bar{W}_{it}$; (i), (2) and (5); (j), $E[Y_{it}|R_{ij} = 1, \bar{W}_{ij}] = E[Y_{it}|R_{i(j-1)} = 1, \bar{W}_{ij}]$, $t \geq j, j = 1, \ldots, T$; and (k), model (j) and Equation (5).

4. DATA MISSING COMPLETELY AT RANDOM

When the data are missing completely at random (i.e., Eq. (2) holds), the weighted least squares estimator of β_0 restricted to the observed outcomes, that is, the solution of

$$n^{-1/2}\sum D_i^*(\beta)\varepsilon_i^*(\beta) = 0, \tag{10}$$

is also consistent and asymptotically normal for estimating β_0 under regularity conditions (Liang and Zeger 1986). Here for each β, $\varepsilon_i^*(\beta) = Y_i^* - g^*(X_i, \beta)$ is equal to the vector of observed residuals and, given $D_i(\beta)$, $D_i^*(\beta)$ is the corresponding submatrix. In this section we compare the asymptotic performance of the estimators that are solutions to equations of the form (7) with the performance of the estimators resulting from (10). Our plan is to find $\hat{\beta}_{comp}$, the best estimator among the solutions to (10), and to provide necessary and sufficient conditions for the asymptotic variance of $\hat{\beta}_{comp}$ to equal the semiparametric variance bound Γ_{opt}^{-1}. Define $U_{comp,i}(\beta) = D_{full,i}^*(\beta)\varepsilon_i^*(\beta)$, where $D_{full,i}(\beta) = \{\partial g(X_i,\beta)/\partial\beta\}\,\mathrm{var}(\varepsilon_i|X_i)^{-1}$. Chamberlain (1987) proved that in the absence of missing data, $\hat{\beta}_{comp}$ solving $\sum_i U_{comp,i}(\beta) = 0$ achieves the semiparametric variance bound in the semiparametric model characterized by (4) even if data on additional time-dependent covariates V_{it} are available. In the presence of missing data, we prove the following lemma in Appendix B.

Lemma 1. Under (2), (4), and (8), (a) $\hat{\beta}_{comp}$ is regular and (b) $n^{1/2}(\hat{\beta}_{comp} - \beta_0)$ has minimum asymptotic variance among all estimators of β_0 that are solutions to equations of the form (10).

Because $\hat{\beta}_{comp}$ is regular, it must have asymptotic variance greater than or equal to that of $\hat{\beta}_{op}$. The following lemma, proved in Appendix B, implies that necessary and sufficient conditions for equality of the asymptotic variance of $\hat{\beta}_{comp}$ and $\hat{\beta}_{op}$ are that the conditional expectations G_{itj} defined in Section 2 do not depend on $\bar{V}_{it}$ and are linear in $\bar{Y}_{it}$ or, equivalently, in $\bar{\varepsilon}_{it}$.

Lemma 2. Suppose that $\bar{\lambda}_{it} < 1$ w.p.1, for $1 \leq t \leq T$. Then the asymptotic variance of $n^{1/2}(\hat{\beta}_{comp} - \beta_0)$ equals Γ_{op}^{-1} if and only if $G_{itj} = G_{itj}^\dagger$ for $j \geq t$, $1 \leq t \leq T$ where $G_{itj}^\dagger = E(\varepsilon_{ij}\bar{\varepsilon}_{it}^T|X_i)\{\mathrm{var}(\bar{\varepsilon}_{it}|X_i)\}^{-1}\bar{\varepsilon}_{it}$.

Define $\hat{\beta}_{lin}$ like $\hat{\beta}_{op}$, except with $G_{itj}^\dagger$ substituted for G_{itj} in the definitions of $\bar{Q}_{it}^*$. The key step in the proof of Lemma 2 is showing that under (2), $\hat{\beta}_{comp}$ and $\hat{\beta}_{lin}$ have the same asymptotic variance. Lemma 2 implies that if data on time-dependent covariates V_{it} correlated with Y_i are available and some data are missing, then under (2), (4), and (8), the optimal solution $\hat{\beta}_{op}$ will have smaller asymptotic variance than $\hat{\beta}_{comp}$ or $\hat{\beta}_{lin}$, because as opposed to $\hat{\beta}_{comp}$ or $\hat{\beta}_{lin}$, $\hat{\beta}_{op}$ exploits the correlation between the past $(V_{i0}^T, V_{i1}^T, \ldots, V_{i(t-1)}^T)$ and Y_{it}. A remarkable property of $\hat{\beta}_{op}$ is that it takes

this correlation into account without making any assumptions about the joint distribution of $(\bar{W}^T_{i(T+1)}, R_{i1}, \ldots, R_{iT})$ other than (2), (4), and (8).

To illustrate these theoretical results, Table 1 gives results for time $t = 4$ of simulation experiments, each based on 200 realizations, so that the estimated coverage probability of a true 95% confidence interval will have a simulation accuracy of approximately ±3% due to Monte Carlo variability. The experiments are similar to those reported in table 2 of Robins et al. (1995), except that missingness is completely at random. Specifically, we conducted two simulation experiments with $X_i = 1$ and $V_{i0} = 1$ and for $t = 1, 2, 3, 4$, $i = 1, \ldots, 500$, $Y_{it} = 200 - 40(t-1) + \sigma_{0i}\{6 - (t-1)\} + \varepsilon_{0it}$, and $V^{1.33}_{it} = 3{,}000 - 100(t-1) + \sigma_{1i}\{10 - (t-1)\} + \varepsilon_{1it}$, 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 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, $\varepsilon_{0i1} \sim N(0, 40^2)$, $\varepsilon_{0i2} \sim N(0, 35^2)$, $\varepsilon_{0i3} \sim N(0, 25^2)$, and $\varepsilon_{0i4} \sim N(0, 10^2)$. Finally, $\bar{\lambda}_{it} = 1$ for $t = 1$ and $\bar{\lambda}_{it} = .6$ for all i and $t = 2, 3, 4$. The data generating and analysis programs were written in Borland's C^{++} version 3.0 using the built-in pseudorandom number generator and were implemented on a Northgate 386 PC. The efficiency of $\hat{\beta}_{\text{comp}}$ and the weighted estimator that uses $\bar{Y}_{it}$ in separate logistic models for the $\bar{\lambda}_{it}$ are equal and exceed that of the sample average estimator $\Sigma_i R_{i4} Y_{i4} / \Sigma_i R_{i4}$. The equal efficiency is predicted by the foregoing theory, because the weighted estimator is asymptotically equivalent to $\hat{\beta}_{\text{lin}}$. However, $\hat{\beta}_{\text{comp}}$ and $\hat{\beta}_{\text{lin}}$ are less efficient than the weighted estimator that uses both $\bar{Y}_{it}$ and $V_{i(t-1)}$ in logistic models for $\bar{\lambda}_{it}$. Finally, even this latter weighted estimator is less efficient than the semiparametric efficient estimator $\hat{\beta}_{\text{op}}$ that uses the optimal covariate $\bar{Q}_{it}$ in the logistic model for $\bar{\lambda}_{it}$. The actual coverage rates of our nominal 95% confidence intervals are as low as 92%. This undercoverage is no longer present at a sample size of 1,000 (data not shown).

Lemma 2 also implies that when only data on Y_{it} and X_i are recorded, $\hat{\beta}_{\text{op}}$ will still have smaller asymptotic variance than $\hat{\beta}_{\text{comp}}$ when the G_{itj} are nonlinear functions of $\bar{Y}_{it}$. But when $Y_{i1}, \ldots, Y_{iT}$ are jointly multivariate normal given X_i and (2) holds, G_{itj} is linear in $\bar{Y}_{it}$, $\hat{\beta}_{\text{op}}$ is $\hat{\beta}_{\text{lin}}$, and both $\hat{\beta}_{\text{comp}}$ and $\hat{\beta}_{\text{op}}$ are asymptotically equivalent to the normal theory maximum likelihood estimator of β_0. Thus knowledge that the data are normally distributed does not asymptotically add information about β_0 when (2), (4), and (8) hold.

5. ADAPTIVE ESTIMATION

$\hat{\beta}_{\text{op}}$ is not available for data analysis, because $\bar{Q}_{it}$ and $D_{\text{op},i}(\beta)$ depend on the unknown probability law generating the data. Our approach will be to replace them by "adaptive" estimates. Specifically, given a preliminary inefficient estimator $\hat{\beta}$ and an estimator $\hat{G}_{itj}$ of G_{itj}, we will estimate, in order, $\bar{Q}^*_{it}$, P^*_i, $D_{\text{op},i}(\beta)$, and $\bar{Q}_{it}$. To estimate the G_{itj}, we cannot simply regress the estimated residuals $\varepsilon_{ij}(\hat{\beta})$ on functions of $\bar{W}_{it}$ among subjects observed at time j, because the missing mechanism (1) does not imply that G_{itj} equals $E(\varepsilon_{ij} | \bar{W}_{it}, R_{ij} = 1)$. But it is straightforward to prove that when (1) holds, $G_{itj} = E(\bar{\pi}_{i(t-1)} \bar{\pi}^{-1}_{ij} \varepsilon_{ij} | R_{ij} = 1, \bar{W}_{it}) P(R_{ij} = 1 | R_{i(t-1)} = 1, \bar{W}_{it})$. Hence we adopt the following two-stage estimation procedure:

Stage 1. Specify flexible regression models

$$P(R_{ij} = 1 | R_{i(t-1)} = 1, \bar{W}_{it}) = l^{(t)}_j(\chi^{(t)}_j, \bar{W}_{it}) \quad (11)$$

and

$$E\{\bar{\pi}_{i(t-1)} \bar{\pi}^{-1}_{ij} \varepsilon_{ij} | R_{ij} = 1, \bar{W}_{it}\} = m^{(t)}_j(\tau^{(t)}_j, \bar{W}_{it}), \quad (12)$$

depending on finite-dimensional parameters $\chi^{(t)}_j$ and $\tau^{(t)}_j$. Often, the right side of (11) will be chosen to be of logistic form. Given $\hat{\alpha}$ and a preliminary inefficient estimate $\hat{\beta}$ of β_0 obtained by solving (7), let $\hat{\chi}^{(t)}_j$ be the maximum likelihood estimator of $\chi^{(t)}_j$ and let $\hat{\tau}^{(t)}_j$ be the (possibly nonlinear) least squares estimator of $\tau^{(t)}_j$ in the regression of $\bar{\pi}_{i(t-1)}(\hat{\alpha}) \varepsilon_{ij}(\hat{\beta}) / \bar{\pi}_{ij}(\hat{\alpha})$ on $\bar{W}_{it}$ among subjects observed at the jth occasion. Then $\hat{G}_{itj} = l^{(t)}_j(\hat{\chi}^{(t)}_j, \bar{W}_{it}) m^{(t)}_j(\hat{\tau}^{(t)}_j, \bar{W}_{it})$, $\hat{\bar{Q}}^*_{itj} = \bar{\pi}_{it}(\hat{\alpha})^{-1} \hat{G}_{itj}$ if $t \le j \le T$ and zero otherwise; $\hat{P}^*_i = \sum_{t=1}^T (R_{it} - \bar{\lambda}^{(\hat{\alpha})}_{it} R_{i(t-1)}) \hat{\bar{Q}}^*_{it}$. We then estimate the $p \times p$ matrix $E\{(U^*_i - P^*_i)^{\otimes 2} | X_i\}$. Given the $(1/2)p(p+1)$ multivariate (possibly nonlinear) regression models

$$E(Z_{ik} Z_{ik'} | X_i) = l_{kk'}(\Psi, X_i) \quad (13)$$

for $1 \le k \le k' \le p$, where Z_{ik} is the kth element of $U^*_i - P^*_i$, estimate Ψ with $\hat{\Psi}$, the (possibly nonlinear) multivariate least squares estimator of the regression of $\hat{Z}_{ik} \hat{Z}_{ik'}$ on X_i. Here $\hat{Z}_{ik}$ is the kth element of $\hat{U}^*_i - \hat{P}^*_i$ and $\hat{U}^*_i = \Delta_i(\hat{\alpha}) \varepsilon_i(\hat{\beta})$. The estimate of $D_{\text{op},i}(\beta)$ is given by $\hat{D}_{\text{op},i}(\beta) = \{\partial g(X_i, \beta)/\partial \beta\} \hat{\text{var}}\{U^*_i - P^*_i | X_i\}^{-1}$, where $\hat{\text{var}}\{U^*_i - P^*_i | X_i\}$ is the $p \times p$ symmetric matrix with (k, k') element $l_{kk'}(\hat{\Psi}, X_i)$. The estimate of $\bar{Q}_{it}$ is given by $\hat{\bar{Q}}_{it} = \hat{D}_{\text{op},i}(\hat{\beta}) \hat{\bar{Q}}^*_{it}$. Notice that by restricting the functions $l_{kk'}(\cdot, X_i)$ such that the symmetric matrix with (k, k') element

Table 1. Results of a Simulation Study at $t = 4$ with $\beta_{0,4} = E(Y_{i4}) = 80.0$

Row	Analysis method	Missingness model	Monte Carlo average of $\hat{\beta}_4$		Actual coverage rate of $\hat{\beta}_4 \pm 1.96\sqrt{\text{s.e.}\hat{\beta}_4}$ in %		Monte Carlo variance of $\hat{\beta}_4$	
			$\rho^2 = .81$	$\rho^2 = .36$	$\rho^2 = .81$	$\rho^2 = .36$	$\rho^2 = .81$	$\rho^2 = .36$
1	Sample average		80.2	80.2	96	94.5	2.4	2.4
2	Optimal GEE ($\hat{\beta}_{\text{comp}}$)		79.9	79.9	94	94.5	2.0	2.0
3	Weighted	$\bar{Y}_{it}$	79.9	80.0	93	93	2.0	2.0
4	Weighted	$\bar{Y}_{it}$, $V_{i(t-1)}$	79.9	80.0	92	92.5	1.8	2.0
5	Weighted, $\hat{\beta}_{\text{op}}$	Optimal	79.9	80.0	95	94.5	1.6	1.8
6	Weighted, $\hat{\beta}_{\text{adap}}$	Adaptive optimal	79.9	80.0	93	93	1.6	1.8

$l_{kk'}(\Psi, X_i)$ is positive definite for all values of Ψ, we guarantee the positive definiteness of $\hat{\text{var}}\{U_i^* - P_i^* \mid X_i\}$.

Stage 2. Obtain $\hat{\omega}$, the partial maximum likelihood of ω_0 in the model (9), where $\bar{Q}_{it}$ is replaced by $\hat{\bar{Q}}_{it}$, and then obtain $\hat{\beta}_{\text{adap}}$, the solution of (7) that uses $\hat{D}_{\text{op},i}(\beta)$ and $\bar{\lambda}_{it}^{(\hat{Q})}(\tilde{\omega})$.

It is standard to show that when (11), (12), and (13) are correctly specified, $\hat{\beta}_{\text{adap}}$ has the same asymptotic distribution as $\hat{\beta}_{\text{op}}$ (see, for example, Robins, Mark, and Newey 1992). Furthermore, $\hat{\beta}_{\text{adap}}$ will be asymptotically unbiased for β_0 even when (11), (12), or (13) are misspecified or even incompatible in the sense that there exists no joint distribution for the observable random variables compatible with (2), (4), and the models (11)–(13). As was shown by Robins et al. (1995, sec. 3), a consistent estimate of the asymptotic variance of $n^{1/2}(\hat{\beta}_{\text{adap}} - \beta_0)$ that is robust to misspecification of (11), (12), or (13) is given by

$$n^{-1}\left\{\sum_{i=1}^{n} \hat{D}_{\text{op}}(\hat{\beta})\partial g(X_i, \hat{\beta})/\partial\beta^T\right\}^{-1} \times \sum_{i=1}^{n}\left\{\hat{U}_i - \left(\sum_{i=1}^{n} \hat{U}_i\hat{S}_{\omega,i}^T\right)\left(\sum_{i=1}^{n} \hat{S}_{\omega,i}\hat{S}_{\omega,i}^T\right)^{-1}\hat{S}_{\omega,i}\right\}^{\otimes 2} \times \left\{\sum_{i=1}^{n} \hat{D}_{\text{op}}(\hat{\beta})\partial g(X_i, \hat{\beta})/\partial\beta^T\right\}^{T^{-1}},$$

where $\hat{U}_i = \hat{D}_{\text{op}}(\hat{\beta})\Delta_j(\hat{\alpha})\varepsilon_i(\hat{\beta})$ and $\hat{S}_{\omega,i} = \sum_{t=1}^{T}(R_{it} - \bar{\lambda}_{it}(\hat{\alpha})R_{i(t-1)})\{\partial \text{ logit } \bar{\lambda}_{it}^{(\hat{Q})}(\tilde{\omega}^*)/\partial\omega\}$ with $\tilde{\omega}^* = (\hat{\alpha}^T, 0^T)^T$. The one-step estimator $\hat{\beta}_{1\text{step}} \equiv \hat{\beta} - \{\sum_{i=1}^{n} \hat{D}_{\text{op}}(\hat{\beta})\partial g(X_i, \hat{\beta})/\partial\beta^T\}^{-1}\sum_{i=1}^{n}\{\hat{U}_i - (\sum_{i=1}^{n} \hat{U}_i\hat{S}_{\omega,i}^T)(\sum_{i=1}^{n} \hat{S}_{\omega,i}\hat{S}_{\omega,i}^T)^{-1}\hat{S}_{\omega,i}\}$ is asymptotically equivalent to $\hat{\beta}_{\text{adap}}$ and avoids the additional calculations involved in Stage 2.

Row 6 of Table 1 summarizes the performance of $\hat{\beta}_{\text{adap}}$ in the simulation experiment of Section 4. $\hat{\beta}_{\text{adap}}$ is seen to be as efficient as $\hat{\beta}_{\text{op}}$, reflecting the fact that the models (11) and (12) were correctly specified. Specifically, we used $\tau_j^{(t)}\dot{\bar{W}}_{it}$ in (12) and $\{1 + \exp(-\chi_j^{(t)})\}^{-1}$ in (11), where $\dot{\bar{W}}_{it}$ is $\bar{W}_{it}$ with V_{it} replaced by $V_{it}^{1.33}$. Note that our model (11) reflected prior knowledge that missingness was completely at random. Further, because X_i was constant, there was no need to fit model (13).

6. EXTENSION TO ARBITRARY MISSING DATA PATTERNS

The assumption that the missing data pattern is monotone is quite restrictive. To allow for arbitrary missing data patterns, following Robins et al. (1995), set $R_{it}^* = R_{it}$ so that $R_i^* = (R_{i1}^*, \ldots, R_{iT}^*)^T$ is the vector of missing visit indicators and 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. Redefine R_{it} so that $R_{it} = 1$ if $R_{i1}^* = R_{i2}^* = \cdots = R_{it}^* = 1$ and $R_{it} = 0$ otherwise. R_{it} is now an artificial "censoring indicator" that is zero once a subject fails to return at any occasion t', for $t' \le t$. Hence $R_{iT} = 1 \Leftrightarrow R_i^* = \mathbf{1}$, where $\mathbf{1}$ is the T vector of 1s. Set $\bar{W}_{i(t+1)}^*(r) = (W_{i0}, r_1W_{i1}, \ldots, r_tW_{it})$ and $\bar{W}_{it}^* = \bar{W}_{it}^*(R_i^*)$. Then $M_i^* = (R_i^{*T}, \bar{W}_{i(T+1)}^{*T})^T$ are the observed data for subject i. Let M_i be subject i's data 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 $R_{i(t-1)} - R_{it} = 1$. The following theorem states that having access to data M_i^* rather than the "monotone" data M_i provides no additional information concerning β_0 if we impose no additional assumptions concerning the missing data mechanism beyond those of Section 2, and if some dropouts do not later return in the sense that

$$P(R_{it'}^* = 0 \mid R_{i(t-1)} - R_{it} = 1, \bar{W}_{it}) > c > 0 \text{ w.p.1} \quad \text{for} \quad T \ge t' > t \quad \text{and some } c. \tag{14}$$

Theorem 3. In the model characterized by (1), (4), (5), and (8), if (14) is true, then the semiparametric variance bound for β_0 based on the data M_i^*, $i = 1, \ldots, n$, equals that based on the data M_i, $i = 1, \ldots, n$.

Suppose that, following Robins et al. (1995), we impose the additional assumption

$$P(R_{it}^* = 1 \mid \bar{R}_{it}^*, \bar{W}_{it}^*, \bar{W}_{i(T+1)}) = P(R_{it}^* = 1 \mid \bar{R}_{it}^*, \bar{W}_{it}^*). \tag{15}$$

Equation (15) implies that the data are missing at random in the sense of Rubin (1976). Redefine $\bar{\lambda}_{it}$ to be $P(R_{it}^* = 1 \mid \bar{R}_{it}^*, \bar{W}_{it}^*)$ 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 (5). Let $\hat{\alpha}$ solve $\Sigma_i S_{\alpha,i}(\alpha) = 0$, where $S_{\alpha,i}(\alpha) = \sum_{t=1}^{T}\{R_{it}^* - \bar{\lambda}_{it}(\alpha)\}\partial \text{ logit } \bar{\lambda}_{it}(\alpha)/\partial\alpha$. Let $\phi_i = \{\phi_{i,r};\ r \ne \mathbf{1}\}$, where $\phi_{i,r}$ is for each $r \ne \mathbf{1}$ a vector-valued function of $\bar{W}_{i(T+1)}(r)$ of dimension ν selected by the investigator. Define $\bar{\pi}_i(r, \alpha) = \prod_{t=1}^{T}\lambda_t(\alpha, \bar{r}_t, \bar{W}_{it}^*(r))^{r_t}[1 - \lambda_t(\alpha, \bar{r}_t, \bar{W}_{it}^*(r))]^{1-r_t}$ and $A_i(\phi, \alpha) = R_{iT}\bar{\pi}_{iT}(\alpha)^{-1}\Sigma_{r\ne\mathbf{1}}\bar{\pi}_i(r, \alpha)\phi_{i,r} - \Sigma_{r\ne\mathbf{1}}I(R_i^* = r)\phi_{i,r}$, where $\bar{r}_t = (r_1, \ldots, r_t)^T$. Following Robins et al. (1995), let $\tilde{\beta}$ solve

$$0 = \sum_i U_i(\beta, \hat{\alpha}) - \hat{\theta}A_i(\phi, \hat{\alpha}), \tag{16}$$

where $\hat{\theta} = \hat{\theta}_1\hat{\theta}_2^{-1}$, $\hat{\theta}_1 = n^{-1}\Sigma_i\hat{\text{Resid}}[U_i(\hat{\beta}, \hat{\alpha}), S_{\alpha,i}(\hat{\alpha})]\hat{\text{Resid}}[A_i(\phi, \hat{\alpha}), S_{\alpha,i}(\hat{\alpha})]^T$, $\hat{\theta}_2 = n^{-1}\Sigma_i\{\hat{\text{Resid}}[A_i(\phi, \hat{\alpha}), S_{\alpha,i}(\hat{\alpha})]\}^{\otimes 2}$, and for any random vectors $\mathbf{A}_i$, $\mathbf{B}_i$, $\hat{\text{Resid}}(A_i, B_i)$ is the residual for subject i from the least squares regression of A_i on B_i, $i = 1, \ldots, n$. In Appendix C we prove the following theorem.

Theorem 4. In the model characterized by (15), (4), (5), and (8), there exist $D_{\text{op},i}^{\dagger}(\beta) = d_{\text{op}}^{\dagger}(X_i, \beta)$ and $\phi_{\text{op},i}$ such that the estimator $\tilde{\beta}$ that solves Equation (16) using these quantities attains the semiparametric variance bound based on the data M_i^*, $i = 1, \ldots, n$.

Remark. The bound is unchanged if we replace Equations (5) and (15) by the weaker assumption that the data are missing at random.

We show in Appendix C that, in contrast to $D_{\text{op},i}(\beta)$, $D_{\text{op},i}^{\dagger}(\beta)$ and $\phi_{\text{op},i}$ depend on the solution to an integral equation that does not exist in closed form, and thus adaptive estimation of these quantities would be computationally complex requiring numerically solving the integral equations by successive approximations (Robins, Rotnitzky and Zhao 1994, sec. 7).

7. ADDITIONAL CONSIDERATIONS

We have assumed that loss to follow-up (i.e., censoring) and measurements of the outcome Y_{it} occur only at discrete

times $t = 1, \ldots, T$. When loss to follow-up and measurements of outcome occur in continuous time, results of Robins and Rotnitzky (1992) and Robins (1995) can still be used both to derive the semiparametric variance bound and to construct a class of estimators which includes an estimator whose asymptotic variance attains the bound. In fact, given any semiparametric model in which the data are missing at random and the probability of observing complete data is bounded away from zero, Robins and Rotnitzky (1992) and Robins (1995) derived both a functional (integral) equation for the efficient score and showed how to construct a class of regular asymptotically linear estimators that includes an efficient estimator. Robins and Rotnitzky (1992) and Robins (1993, 1995) constructed such a class of estimators in the accelerated failure time model, a median regression failure time model, the one-sample survival model and the Cox proportional hazards regression model in the presence of dependent censoring attributable to time-dependent covariates that simultaneously predict failure and censoring. Robins et al. (1994) and Pugh, Robins, Lipsitz and Harrington (1993) constructed such a class of estimators in a conditional mean model and the Cox proportional hazards model in the presence of missing covariates respectively. In contrast to the results obtained in this article for the conditional mean model with missing outcome Y, Robins et al. (1994) showed that even with a monotone missing data pattern, there is no closed-form expression for the efficient score in a conditional mean model with missing covariates X when W_{it} has continuous components.

Robins et al. (1994, sec. 7.1) showed that with monotone missing data, the class of estimators $\hat{\beta}^*$ solving $\sum_{i=1}^n U_i^*(\beta, \hat{\alpha}) = 0$ with $U_i^*(\beta, \alpha) = \bar{\pi}_{iT}(\alpha)^{-1} R_{iT} D_i(\beta) \varepsilon_i(\beta)$ also contains an estimator whose asymptotic variance attains the semiparametric efficiency bound for our model. But because the $U_i^*(\beta, \hat{\alpha})$ are nonzero only if subject i has complete data, the small-sample behavior of the estimators $\hat{\beta}^*$ are inferior to those of the estimators $\hat{\beta}$ whenever $\bar{\pi}_{iT}(\alpha_0)$ can be small, say less than .2 or .3. It is for this reason that we have restricted attention to the estimators $\hat{\beta}$.

The estimator $\hat{\beta}_{\text{adap}}$ of Section 5 is locally semiparametric efficient in the model characterized by (2), (4) and (8) because it attains the variance bound if models (11)–(13) are true and remains regular, asymptotically linear even if they are false (Bickel et al., 1993). The good performance of $\hat{\beta}_{\text{adap}}$ in the moderate sized samples of our simulation study depends critically X being of small dimension. Specifically, suppose that the missing data pattern is monotone, the regression vector X_i is multivariate with a large number of continuous components (say, 5), that we are willing to assume that, given X_i, the data are missing completely at random (i.e., Eq. (2) holds), but that we do not wish to specify any model for $P[R_{it} = 1 \mid R_{i(t-1)} = 1, X_i]$ whatsoever. Hence our semiparametric model is again characterized by (2), (8) and the conditional mean restriction (4). Then in general, any inverse probability of censoring weighted estimator $\hat{\beta}$ (such as $\hat{\beta}_{\text{adap}}$) that is more efficient than $\hat{\beta}_{\text{comp}}$ will perform poorly in moderate size samples (in the sense that there will be distributions allowed by our semiparametric model under which $\hat{\beta}$ fails to be centered at β_0 with an approximately normal sampling distribution). This poor performance is due to the fact that $\hat{\beta}$ will not be centered on β_0 unless we can obtain an accurate estimate of $\bar{\pi}_{it}$; however, due to the curse of dimensionality, it is not possible to obtain an accurate estimate because estimation of the unrestricted conditional expectations given X_i that occur in $\bar{\pi}_{it}$ requires smoothing in at least five dimensions. In contrast, GEE-like estimators solving (10) will perform well in moderate size samples because no high dimensional smooths are needed. In fact, in a separate report, we argue that any regular asymptotically linear estimator in this model with asymptotic variance less than that of $\hat{\beta}_{\text{comp}}$ will require accurate estimation of unrestricted conditional expectations given X_i and thus, due to the curse of dimensionality, will perform poorly in moderate size samples.

APPENDIX A: PROOF OF THEOREMS 1 AND 2

Our proof of Theorem 1 uses a general representation for the semiparametric efficient score of an arbitrary semiparametric model with monotone missing at random data given by Robins and Rotnitzky (1992, thm. 4.2) and Robins et al. (1994, props. 8.1–8.2).

We start with a review of the theory of semiparametric efficiency bounds, borrowing heavily from the survey paper of Newey (1990) and the monograph of Bickel et al. (1992). Using the notation in the first paragraph of Section 3, define the nuisance tangent space Λ to be the closed linear span of all random vectors bS_η, where S_η is the subject-specific score for η evaluated at the truth in some regular parametric submodel, usually $S_\eta = \partial \ln L(\beta_0, \eta_0; Z)/\partial\eta$; b is a conformable constant matrix with p rows; and the subscript i has been suppressed. Here consider Λ as a subset of the Hilbert space of $p \times 1$ random vectors H with inner product $E(H_1^T H_2)$ and $E(H^T H) < \infty$. Then the projection of any vector H on Λ exists and is the unique vector $\Pi(H|\Lambda)$ in Λ satisfying $E[\{H - \Pi(H|\Lambda)\}^T A] = 0$ for all A in Λ. Π is a projection operator.

The semiparametric variance bound for regular estimators of β_0 equals the inverse of the variance of $S_{\text{eff}} = \Pi(S_\beta|\Lambda^\perp)$, where S_β is the score for β in the semiparametric model $L(\beta, \theta; Z)$, usually $S_\beta = \partial \ln L(\beta_0, \theta_0; Z)/\partial\beta$, and $\Lambda^\perp$ is the orthogonal complement of Λ. The random variable S_{eff} is called the efficient score. Further, any regular, asymptotically linear estimator $\hat{\beta}$ with asymptotic variance $\{\text{var}(S_{\text{eff}})\}^{-1}$ has the efficient influence function $\{\text{var}(S_{\text{eff}})\}^{-1} S_{\text{eff}}$, where $\hat{\beta}$ is asymptotically linear with influence function U if $n^{1/2}(\hat{\beta} - \beta_0) = \Sigma_i U_i + o_p(1)$, $E(U) = 0$, and $\text{var}(U^T U) < \infty$. We now specialize the foregoing general results to the "full" and "missing" data models of Section 2.

Let $Z^{(F)} = \bar{W}_{(T+1)}$ be the data vector that would be available on a subject in the absence of missing data. Let $L^{(F)}(\beta, \theta; Z^{(F)})$ be the likelihood for a single subject when $Z^{(F)}$ is fully observed in the full-data semiparametric model characterized by (4), indexed by the $p \times 1$ vector β and an infinite-dimensional nuisance parameter θ and let $S_\beta^{(F)}$, $\Lambda^{(F)}$, and $S_{\text{eff}}^{(F)}$ be the score for β, the nuisance tangent space, and the efficient score for the full data model. Let Z be the observed data vector for a subject in the presence of monotone missing data, so that $Z = (\bar{W}_t^T, \bar{R}_{(T+1)}^T)^T$ if the subject was lost to follow-up at time t and $Z = (Z^{(F)}, \bar{R}_{(T+1)})$ in the absence of missing data. Let $L(\beta, \theta, Z)$ be the likelihood for a single subject in the semiparametric model characterized by (3) and (4). For any set $\mathcal{F}$ of random variables, let $\mathcal{F}_0$ be the subset of variables in $\mathcal{F}$ with mean zero. Our proof of Theorem 1 uses the following two lemmas, the first of which is a restatement of proposition (8.3) of Robins et al. (1994).

Lemma A.1. Under the full-data semiparametric model characterized by the sole restriction Equation (4), (a) $S_{\text{eff}}^{(F)} = E(S_\beta^{(F)}\varepsilon^T|X)E(\varepsilon\varepsilon^T|X)^{-1}\varepsilon$ with $E(S_\beta^{(F)}\varepsilon^T|X) = \partial g(X, \beta_0)/\partial\beta$; (b) If $E(H) = 0$ for $H = h(Z^{(F)})$, $\Pi(H|\Lambda^{(F)\perp}) = E(H\varepsilon^T|X)E(\varepsilon\varepsilon^T|X)^{-1}\varepsilon$; and (c) $\Lambda_0^{(F),\perp} = \{d(X)\varepsilon\}$, where $\Lambda_0^{(F),\perp} \equiv (\Lambda^{(F),\perp})_0$.

Lemma A.2. The efficient score S_{eff} in a semiparametric model $L(\beta, \theta; Z)$ with monotone missing data satisfying the restrictions (3) and (8) satisfies

$$S_{\text{eff}} = Q_{\text{eff}} + \sum_{t=1}^{T} (R_t - \bar{\lambda}_t R_{t-1})\bar{\pi}_t^{-1}\{Q_{\text{eff}} - E[Q_{\text{eff}}|\bar{W}_t, R_{t-1} = 1]\}, \tag{A.1}$$

where Q_{eff} is the unique Q in $\Lambda^{(F)\perp}$ satisfying

$$S_{\text{eff}}^{(F)} = Q + \Pi[v(Q)|\Lambda^{(F)\perp}], \tag{A.2}$$

with $v(Q) = \sum_{t=1}^{T}(1 - \bar{\lambda}_t)\bar{\pi}_t^{-1}\{Q - E(Q|\bar{W}_t, R_{t-1} = 1)\}$.

Proof of Lemma A.2

This is a specific application of theorems (4.1f) and (4.2) of Robins and Rotnitzky (1992) or of propositions (8.1)–(8.2) of Robins et al. (1994).

Proof of Theorem 1

By Lemma A.1c, $Q_{\text{eff}} = D_{\text{eff}}\varepsilon = d_{\text{eff}}(X)\varepsilon$. Substituting $D_{\text{eff}}\varepsilon$ for Q_{eff} in (A.1), we obtain that $S_{\text{eff}} = D_{\text{eff}}(\Delta\varepsilon - P^*)$, because $\Delta\varepsilon = \varepsilon + \sum_{t=1}^{T}(R_t - \bar{\lambda}_t R_{t-1})\bar{\pi}_t^{-1}I^{(t)}\varepsilon$, where $I^{(t)}$ is the $T \times T$ diagonal matrix with diagonal elements $e_{jj} = 1$ if $j \geq t$ and $e_{jj} = 0$ otherwise. Note that $\Delta\varepsilon - P^* = \varepsilon + M(T)$, where

$$M(t) = \sum_{j=1}^{t}(R_j - \bar{\lambda}_j R_{j-1})\bar{\pi}_j^{-1}\{\varepsilon - E[\varepsilon|\bar{W}_j, R_{j-1} = 1]\}. \tag{A.3}$$

Further, it follows by equation (A.4) in the proof of lemma 1 of Robins et al. (1995) that $n^{1/2}(\hat{\beta}_{\text{op}} - \beta_0) = \Gamma_{\text{op}}^{-1}n^{-1/2}\Sigma_i D_{\text{op},i}(\Delta_i\varepsilon_i - P_i^*) + o_p(1)$, where $D_{\text{op}} = D_{\text{op}}(\beta_0)$. Thus by the central limit theorem, $\text{var}^A\{n^{1/2}(\hat{\beta}_{\text{op}} - \beta_0)\} = \{\text{var}(S_{\text{eff}})\}^{-1}$ if we can show $D_{\text{op}} = D_{\text{eff}}$ and $\Gamma_{\text{op}} = \text{var}(S_{\text{eff}})$. Using Lemma A.1c to write $Q = d(X)\varepsilon$, Equation (A.2) becomes $\{\partial g(X, \beta_0)/\partial\beta\}E(\varepsilon\varepsilon^T|X)^{-1}\varepsilon = d(X)\varepsilon + E[v(Q)\varepsilon^T|X]E(\varepsilon\varepsilon^T|X)^{-1}\varepsilon$. Further, $E[v(Q)\varepsilon^T|X] = E[\sum_{t=1}^{T}(1 - \bar{\lambda}_t)\bar{\pi}_t^{-1}\{d(X)\varepsilon - E[d(X)\varepsilon|\bar{W}_t, R_{t-1} = 1]\}\varepsilon^T|X] = d(X)K(X)$, with $K(X) = E\{\sum_{t=1}^{T}(1 - \bar{\lambda}_t)\bar{\pi}_t^{-1}\text{var}(\varepsilon|\bar{W}_t, R_{t-1} = 1)|X\}$. For (A.2) to be true for all ε, it is necessary that $\{\partial g(X, \beta_0)/\partial\beta\}E(\varepsilon\varepsilon^T|X)^{-1} = d(X) + d(X)K(X)E(\varepsilon\varepsilon^T|X)^{-1}$, which implies that $D_{\text{eff}} = \{\partial g(X, \beta_0)/\partial\beta\}\{E(\varepsilon\varepsilon^T|X) + K(X)\}^{-1}$. To see that $D_{\text{op}} = \{\partial g(X, \beta_0)/\partial\beta\}\text{var}(\Delta\varepsilon - P^*|X)^{-1}$ equals D_{eff}, note that $\Delta\varepsilon - P^* = \varepsilon + M(T)$, where, as shown in the proof of theorem 1 of Robins et al. (1995), $M(t)$ of Equation (A.3) is a discrete time mean zero martingale process with respect to the filtration $\sigma\{\bar{W}_t, \bar{R}_t, \varepsilon\}$, and thus $E[\varepsilon M(T)^T|X] = 0$ and, by a standard martingale variance calculation, $\text{var}[M(T)|X] = K(X)$. A similar calculation shows that $E(S_{\text{eff}}S_{\text{eff}}^T) = \Gamma_{\text{opt}}$.

Proof of Theorem 2

Theorem (4.1f) of Robins and Rotnitzky (1992) and proposition (8.1) of Robins et al. (1994) state that when the data are missing at random (i.e., Eq. (3) is true), the efficient score and semiparametric variance bound do not depend on a model for, or knowledge of, the missingness probabilities $\bar{\lambda}_t$. This implies that the bound in models (e)–(i) equals that in model (a). To show that models (a) and (b) have the same bound, note that model (a) differs from model (b) only by imposing the additional nonidentifiable restriction

$$f(V_t|\bar{W}_t, \varepsilon) = f(V_t|\bar{W}_t, R_t = 1, \varepsilon). \tag{A.4}$$

Restriction (A.4) is nonidentifiable, because it may be true whatever be the distribution of the observables Z. Indeed, models (a), (b), and (j) imply exactly the same restrictions on the distribution of the observables—namely, $g_t(X_i, \beta_0) = \int\int\int\cdots\int E[Y_{it}|\bar{w}_t, R_{it} = 1, X_i]\prod_{j=0}^{j=t-1} dF[w_j|\bar{w}_j, R_{ij} = 1, X_i]$, $t = 1, \ldots, T$, which, by lemma (A.1) of Robins, et al. (1995), is equivalent to the restrictions $g_t(X_i, \beta_0) = E[R_{it}\bar{\pi}_{it}^{-1}Y_{it}|X_i]$. Hence the allowable densities for the observable random variables Z under the models (a), (b), and (j) are the same, and thus by definition the implied semiparametric models for the observables Z are identical. In particular, the semiparametric variance bound for estimating β_0 will be the same functional of the distribution function of the observables. Finally, analogous arguments show that the bounds in models (c), (d) and (e) are identical.

APPENDIX B: PROOF OF LEMMAS 1 AND 2

We prove Lemma 1a under regularity conditions (R.8) and (R.9) of appendix A of Robins et al. (1994). Lemma 1a follows from the fact that (a) $U_{\text{comp}}(\beta)$ has mean zero when $\beta = \beta_0$ under (2), (4), and (8), (b) $U_{\text{comp}}(\beta)$ is continuously differentiable in β_0, and (c) by Theorem 2.2 in Newey (1990), solutions to differentiable unbiased estimating equations are regular under our regularity assumptions. Note that Lemma 1b is true when the distribution of $\varepsilon|X$ is multivariate normal (MVN), because $\hat{\beta}_{\text{comp}}$ is the parametric maximum likelihood estimator (MLE) of β_0 in the model that imposes, in addition to (2), (4), and (8), the assumption that $\varepsilon|X$ is MVN with mean zero and *known* covariance matrix $\text{var}(\varepsilon|X)$. We now show that $\hat{\beta}_{\text{comp}}$ is at least as efficient as $\hat{\beta}_d$ that solves (10) based on $D(\beta) = d(X, \beta)$, even if $\varepsilon|X$ is not MVN. Because $D^*(\beta)\varepsilon^*$ is linear in ε, we obtain, using the usual formula for the asymptotic variance of a solution to an unbiased estimating equation, that the asymptotic variance of $\hat{\beta}_d$ depends on the law of $\varepsilon|X$ only through $\text{var}(\varepsilon|X)$. In particular, the efficiency ordering cannot depend on whether $\varepsilon|X$ is MVN, proving Lemma 1b. Turning now to Lemma 2, because $\hat{\beta}_{\text{comp}}$ is a parametric MLE when $\varepsilon|X$ is MVN and is regular in the semiparametric model (2), (4), and (8), the asymptotic variance of $\hat{\beta}_{\text{comp}}$ must equal that of $\hat{\beta}_{\text{op}}$ when $\varepsilon|X$ is MVN. But when $\varepsilon|X$ is MVN, it is easy to show that $G_{itj} = G_{itj}^\dagger$ and $\hat{\beta}_{\text{lin}} = \hat{\beta}_{\text{op}}$. Now, even when $\varepsilon|X$ is not MVN, the influence function of $\hat{\beta}_{\text{lin}}$, in contrast to that of $\hat{\beta}_{\text{op}}$ remains linear in ε and has an asymptotic variance that only depends on the law of $\varepsilon|X$ through $\text{Var}(\varepsilon|X)$. Hence $\hat{\beta}_{\text{lin}}$ and $\hat{\beta}_{\text{comp}}$ have the same asymptotic variance whatever be the law of $\varepsilon|X$. However $\hat{\beta}_{\text{lin}}$ has asymptotic variance equal to that of $\hat{\beta}_{\text{op}}$ if and only if $G_{itj} = G_{itj}^\dagger$ by the uniqueness of the efficient influence function.

APPENDIX C: PROOF OF THEOREMS 3 AND 4

Proof of Theorem 3

Let C record the first missed visit, so $C = t$ if $R_{t-1} - R_t = 1$ and, by convention, $C = T + 1$ if $R_T = 1$. Set $R_{T+1}^* \equiv 0$. In the model of Theorem 3, the likelihood function $L(\beta, \theta; M^*)$ is given by $\{\int\cdots\int L^{(F)}(\beta, \theta; \bar{W}_{T+1})f(\underline{R}_C^*|C, \bar{W}_{T+1}; \theta)\prod_{m=C}^{T+1} dW_m^{(1-R_m^*)}\}\{(1 - \bar{\lambda}_C)\prod_{m=1}^{C-1}\bar{\lambda}_m\}$, where for any $a = (a_1, \ldots, a_T)^T$, $\underline{a}_t = (a_{t+1}, \ldots, a_T)^T$ and, without loss of generality, we assume that $\bar{\lambda}_m$ is known. Then $S_\beta = E(S_\beta^{(F)}|R^*, W^*)$ with $W^* \equiv \bar{W}_{T+1}^*$ is the score for β based on data M_i^* by proposition (A5.5) of Bickel et al. (1993). Define $S_\beta^{\text{mon}} = E(S_\beta^{(F)}|C, \bar{W}_C)$ and $S_\beta^{\text{resid}} = S_\beta - S_\beta^{\text{mon}}$. S_β^{mon} is the score for β based on the data M_i. Let $\Lambda_1 = \{A_1 = E(A^{(F)}|R^*, W^*); A^{(F)} \in \Lambda^{(F)}\}$ and $\Lambda_1^{\text{mon}} = \{A_1^{\text{mon}}E(A^{(F)}|C, \bar{W}_C); A^{(F)} \in \Lambda^{(F)}\}$, $\Lambda_1^{\text{Resid}} = \{A_1^{\text{Resid}} = A_1 - A_1^{\text{mon}}\}$. Let $\Lambda^{\text{mis}} = \{A^{\text{mis}} = a(\underline{R}_C^*, C, \bar{W}_{T+1}); E(A^{\text{mis}}|C, \bar{W}_{T+1})$

$= 0\}$ and let $\Lambda_2 = \{A_2 = E(A^{\text{mis}}|R^*, W^*); A^{\text{mis}} \in \Lambda^{\text{mis}}\}$. Then it is straightforward to show that the nuisance tangent spaces are $\Lambda = \Lambda_1 \oplus \Lambda_2 = \Lambda_1^{\text{Resid}} \oplus \Lambda_1^{\text{mon}} \oplus \Lambda_2$ for the model based on the data M_i^* and Λ_1^{mon} for the model based on the data $M_i = (C_i, \bar{W}_{iC_i})$. Note that $E(A_2|C, \bar{W}_C) = 0$, by definition of A^{mis}. We wish to show that the efficient score $S_{\text{eff}} = \Pi(S_\beta|\Lambda^\perp)$ based on data M_i^*, $i = 1, \ldots, n$, is the same as the efficient score $S_{\text{eff}}^{\text{mon}} = \Pi(S_\beta^{\text{mon}}|\Lambda_1^{\text{mon},\perp})$ based on data M_i, $i = 1, \ldots, n$. Now it is straightforward to check that $S_\beta^{\text{mon}} \perp (\Lambda_1^{\text{resid}} \oplus \Lambda_2)$. Therefore, $\Pi(S_\beta|\Lambda^\perp) = \Pi(S_\beta^{\text{mon}}|\Lambda_1^{\text{mon},\perp}) + \Pi(S_\beta^{\text{Resid}}|\Lambda^\perp)$. The following lemma shows that $S_\beta^{\text{Resid}} \in \Lambda_2$ and hence $\Pi(S_\beta^{\text{Resid}}|\Lambda^\perp) = 0$.

Lemma. $S_\beta^{\text{Resid}} = \mathring{A}_2$ when $\mathring{A}_2 = E(\mathring{A}^{\text{mis}}|R^*, W^*)$ with $\mathring{A}^{\text{mis}} = -\{I(\underline{R}_C^* = \underline{0}_C)\pi(\underline{0}_C)^{-1}\{1 - \pi(\underline{0}_C)\} - \Sigma_{\underline{r}_C \neq \underline{0}_C} I(\underline{R}_C^* = \underline{r}_C)\}\{S_\beta^{(F)} - S_\beta^{\text{mon}}\}$, where $\underline{r}_C = (r_{C+1}, \ldots, r_T)$, $\underline{0}_C = (0_{C+1}, \ldots, 0_T)$, and $\pi(\underline{r}_C) \equiv P(\underline{R}_C^* = \underline{r}_C|C, \bar{W}_{T+1})$.

Proof. Clearly, $\mathring{A}_2$ is in Λ_2 by Equation (14) and $E(\mathring{A}^{\text{mis}}|C, \bar{W}_{T+1}) = 0$. Further, by construction, if $\underline{R}_C^* \neq \underline{0}_C$, then $\mathring{A}_2 = S_\beta^{\text{Resid}}$. Hence it suffices to show that $E(S_\beta^{(F)} - S_\beta^{\text{mon}} - \mathring{A}^{\text{mis}}|\underline{R}_C^* = \underline{0}_C, \underline{W}_C^*, C, \bar{W}_C) = 0$. But this evaluates to $E[\{\pi(\underline{0}_C)\}^{-1}(S_\beta^{(F)} - S_\beta^{\text{mon}})|\underline{R}_C^* = \underline{0}_C, \underline{W}_C^*, C, \bar{W}_C]$, which equals $\{P(\underline{R}_C^* = \underline{0}_C|C, \bar{W}_C)\}^{-1}E(S_\beta^{(F)} - S_\beta^{\text{mon}}|C, \bar{W}_C) = 0$.

Proof of Theorem 4

For notational convenience, define $A(\phi) = A(\phi, \alpha_0)$, $W^*(r) = \bar{W}_{T+1}^*(r)$, $W^* = W^*(R^*) = \bar{W}_{T+1}^*$, $\pi(r) = \bar{\pi}(r, \alpha_0)$, and $R_{T+1}^* = 0$. Let $\Lambda^{(2)} = \{A^{(2)} = a^{(2)}(R^*, W^*); E(A^{(2)}|\bar{W}_{T+1}) = 0\}$ be the set of functions of the observed data $M^\dagger$ that have mean zero given the full data $\bar{W}_{T+1}$. Robins and Rotnitzky (1992) proved that $\Lambda^{(2)}$ is closed. Let $\Pi(\cdot|\Lambda^{(2)})$ be the Hilbert space projection on to $\Lambda^{(2)}$. According to theorem (4.1f) of Robins and Rotnitzky (1992) or proposition (8.1) of Robins et al. (1994), because Equation (15) implies that the data are missing at random in the sense of Rubin (1976), the efficient score in the semiparametric model for Theorem 4 is $S_{\text{eff}} = R_T\bar{\pi}\bar{T}^{-1}Q_{\text{eff}}^\dagger - \Pi(R_T\bar{\pi}\bar{T}^{-1}Q_{\text{eff}}^\dagger|\Lambda^{(2)})$, where

$$Q_{\text{eff}}^\dagger = m(B^\dagger), \tag{A.5}$$

$m(B) \equiv \Sigma_r \pi(r)E[B|W^*(r)]$, and $B^\dagger$ is the unique function of $\bar{W}_{T+1}$ satisfying

$$m(B^\dagger) \in \Lambda_0^{(F),\perp} \quad \text{and} \quad \Pi(B^\dagger|\Lambda^{(F),\perp}) = S_{\text{eff}}^{(F)}. \tag{A.6}$$

Furthermore, according to theorem (4.1g) of Robins and Rotnitzky (1992) or proposition (8.1) of Robins et al. (1994), $\Pi(R_T\bar{\pi}\bar{T}^{-1}Q_{\text{eff}}^\dagger|\Lambda^{(2)})\ R_T\bar{\pi}\bar{T}^{-1}Q_{\text{eff}}^\dagger - \Sigma_r I(R = r)B^\dagger$, which, by (A.5), equals $A(\phi_{\text{eff}})$ with $\phi_{\text{eff},r} = E(B^\dagger|W^*(r))$. Hence

$$S_{\text{eff}} = R_T\bar{\pi}\bar{T}^{-1}Q_{\text{eff}}^\dagger - A(\phi_{\text{eff}}). \tag{A.7}$$

Further, by Lemma A.1, $Q_{\text{eff}}^\dagger = m(B^\dagger) \in \Lambda_0^{(F),\perp}$ implies that $Q_{\text{eff}}^\dagger = D_{\text{eff}}^\dagger\varepsilon$ for $D_{\text{eff}}^\dagger = (D_{\text{eff},1}^\dagger, \ldots, D_{\text{eff},T}^\dagger)$ satisfying $D_{\text{eff},j}^\dagger = E[m(B^\dagger)|Y = e_j, X] - E[m(B^\dagger)|Y = 0, X]$, with e_j the T vector with jth component 1 and the remaining component zero.

Now let $U_{\text{op}}(\beta, \alpha)$ be $U(\beta, \alpha)$ that uses $D(\beta) = D_{\text{eff}}$ for each β. Let $U_{\text{op}} = U_{\text{op}}(\beta_0, \alpha_0)$. Then straightforward algebra shows that $U_{\text{op}} = R_T\bar{\pi}\bar{T}^{-1}D_{\text{eff}}^\dagger\varepsilon - A(\phi^*)$ with $\phi_r^* \equiv D_{\text{eff}}^\dagger(e_1/\bar{\pi}_1, \ldots, e_{t^*(r)}/\bar{\pi}_{t^*(r)}, 0, 0, \ldots, 0)^T$ and $t^*(r) = t$ if $r_1 = \cdots = r_t = 1$, and $r_{t+1} = 0$, $1 \leq t \leq T$. It follows from (A.7) that $S_{\text{eff}} = U_{\text{op}} - A(\phi_{\text{op}})$, where $\phi_{\text{op},r} = \phi_{\text{eff},r} - \phi_r^*$. Further, $\theta_{\text{op}} = E[U_{\text{op}}A(\phi_{\text{op}})^T]\{E[A(\phi_{\text{op}})A(\phi_{\text{op}})^T]\}^{-1}$ is the identity matrix, becasuse $\Pi(U_{\text{op}}|\Lambda^{(2)}) = A(\phi_{\text{op}})$.

The asymptotic variance to the solution to $\Sigma_i U_{\text{op},i}(\beta, \alpha_0) - \theta_{\text{op}}A_i(\phi_{\text{op}}, \alpha_0)$ will attain the semiparametric variance bound, because estimators based on the efficient score attain the bound (Newey 1990). Finally, the asymptotic variance of $\hat{\beta}^\dagger$ solving $\Sigma_i U_{\text{op},i}(\beta, \hat{\alpha}) - \hat{\theta}A_i(\phi_{\text{op}}, \hat{\alpha})$ will attain the bound, because Robins et al. (1994) showed in the proof of their Theorem 2 that replacing θ_{op} by its estimate $\hat{\theta}$ and α_0 by its estimate $\hat{\alpha}$ can never increase the variance of the solution. This proves Theorem 4 with $D_{\text{op}}^\dagger(\beta) = D_{\text{eff}}^\dagger$.

We have provided closed-form expressions for ϕ_{op} and $D_{\text{op}}^\dagger(\beta) = D_{\text{eff}}^\dagger$ in terms of $B^\dagger$. But $B^\dagger$ itself solves an integral equation whose solution does not exist in closed form. Specifically, Robins et al. (1994, sec. 7.2) showed that (A.6) implies $B^\dagger$ solves their equation (47) with X^* instead of X.

[Received April 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.

Bickel, P., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Inference in Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Carroll, R. J., and Ruppert, D. (1982), "Robust Estimation in Heteroscedastic Linear Models," *The Annals of Statistics*, 10, 429–441.

Chamberlain, G. (1987), "Asymptotic Efficiency in Estimation With Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–324.

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

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

McCullagh, P. (1983), "Quasi-Likelihood Functions," *The Annals of Statistics*, 11, 59–67.

Newey, W. K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

——— (1991), "Efficient Estimation of Models With Conditional Moment Restrictions," working paper, Massachusetts Institute of Technology.

Pugh, M., Robins, J. M., Lipsitz, S., and Harrington, D. (1993), "Inference in the Cox Proportional Hazards Model With Missing Covariates," technical report, Harvard School of Public Health, Dept. of Biostatistics.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications* (2nd ed.), New York: John Wiley.

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

——— (1995), "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., Mark, S. D., and Newey, W. K. (1992), "Estimating Exposure Effects by Modeling the Expectation of Exposure Conditional on Confounders," *Biometrics*, 48, 479–495.

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.

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, 846–866.

———, (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes Under the Presence of Missing Data," *Journal of the American Statistical Association*, 90, 106–121.

Robinson, P. (1987), "Asymptotically Efficient Estimation in the Presence of Heteroscedasticity of Unknown Form," *Econometrica*, 55, 875–891.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

Weddeburn, R. W. M. (1976), "On the Existence and Uniqueness for Certain Generalized Linear Models," *Biometrika*, 63, 27–32.