

Semi-parametric Estimation of Models for Means and Covariances in the Presence of Missing Data

ANDREA ROTNITZKY and JAMES M. ROBINS

Harvard School of Public Health

ABSTRACT. In this article we describe a class of inverse-probability-of-censoring-weighted estimating equations for jointly estimating the parameters of models for the conditional mean and covariance of a vector of responses given a set of regressors in the presence of monotone missing outcome data. Our methods are valid when the data are missing at random in the sense of Rubin (1976) and do not require a parametric model for the joint distribution of the data. However, they do require a model for the non-responsive probabilities. We show that the solution to the optimal estimating equation in our class has asymptotic variance equal to the semiparametric variance bound. Because the optimal estimating equation depends on unknown population parameters, we propose an adaptive locally efficient estimator whose asymptotic variance can achieve the semiparametric variance bound.

Key words: adaptive estimation, generalized estimating equations, missing at random, semi-parametric efficiency bound

1. Introduction

Gourieroux *et al.* (1984), Prentice (1988), Zhao & Prentice (1990) and Prentice & Zhao (1991) proposed estimating equations for jointly estimating the parameters β_0 indexing a model for the conditional mean and covariance of a vector of responses Y given a set of regressors X . As noted by these authors, with missing data on Y their approach requires the assumption that the data are missing completely at random (Rubin, 1976), i.e. that, given the covariates, the probability of non-response is independent of observed and unobserved data. The goal of this paper is to provide methods for estimating β_0 when the non-response patterns are monotone and the data are missing at random (Rubin, 1976), i.e. when the probability of non-response may depend on the observed data and is independent of the missing data given the observed data.

One approach to inferences about β_0 in the presence of missing at random data is to specify a fully parametric model for the law of Y given X and to estimate β_0 by maximum likelihood ignoring the missing data process. Fully parametric inferences, however, can be non-robust to misspecification of the full-data likelihood since they implicitly work by imputing the missing data from their conditional distribution given the observed data. Another approach is the pseudo-likelihood method of Arminger & Sobel (1990). This approach does not require specification of the joint distribution of the full data but, contrary to the claim of the authors, it can result in inconsistent estimates of β_0 when the data are missing at random but not completely at random. In this paper we describe an estimation procedure that assumes a model for the conditional means and covariances, but does not require the complete specification of the distribution of the data.

Models for the covariances of the responses can be the focus of scientific interest or can be used to improve the efficiency in the estimation of mean parameters. As an example of the former, consider the longitudinal study reported by DeGruttola *et al.* (1991) of the progression of infection with the human immunodeficiency virus (HIV) where serial measurements of T-helper lymphocytes (T4) were available on a cohort of HIV-infected and uninfected individuals. The goal of the study was to model the evolution of T4 counts over

time, as well as to determine the variability of the individual rates of change of T4 counts in the infected and uninfected cohorts. Their model for the variability of the individual rates implied parametric restrictions on the form of the covariance matrix of the responses. In their study, subjects who developed AIDS under follow-up were more likely to drop out. The methods described in this paper are useful under this non-response scenario.

Models for means and covariances are useful even when the scientific focus is on the description of the mean of the responses only. Chamberlain (1987) showed that the semi-parametric information bound (as defined by Begun *et al.*, 1983) in the semi-parametric model defined only by restrictions on the mean of the outcomes given the covariates is less than or equal, and typically strictly less than, the semi-parametric information bound in the semi-parametric model defined by restrictions on both the means and covariances. Thus, typically, knowledge of the model for the covariance of the repeated outcomes given the covariates helps in estimating the parameters of the model for the means.

In this paper we propose a class of semi-parametric estimators of β_0 that are consistent when the non-response patterns are monotone and the data are missing at random provided one can specify a model for the non-response probabilities. Our class contains a member whose asymptotic variance attains the semi-parametric variance bound in the model defined by restrictions on the conditional means and covariances of the responses given a set of regressors and the condition that the data are missing at random. The focus is the setting of longitudinal studies, in which monotone non-response patterns are common. However, our results are not restricted to longitudinal studies and are applicable whenever the pattern of missing response is monotone.

In section 2 we present our model. In section 3 we introduce a class of estimating equations and we show that it contains the efficient score equations. Because the optimal equations depend on unknown population parameters they are not useful for data analysis. Thus, in section 4 we propose an adaptive estimator whose asymptotic variance can achieve the semiparametric variance bound. Section 5 contains some final remarks.

2. The model

We consider a follow-up study conducted over a fixed interval from time 1 to T . Let $Y_i = (Y_{i0}, Y_{i1}, \dots, Y_{iT})^T$ be the vector of outcome variables corresponding to the i th subject, $i = 1, \dots, n$, measured at prespecified visit times $(0, 1, 2, \dots, T)$ where visit 0 occurs at a time just prior to start of follow-up. Here and throughout T , when used as a superscript, denotes matrix transposition. Let X_i be a vector of baseline explanatory variables measured prior to start of follow-up. We shall assume the following regression models for the first two moments of the outcomes:

$$E(Y_{it} | X_i) = q_t(X_i; \beta_0) \quad (2.1)$$

and

$$E(Y_{it} Y_{it'} | X_i) = h_{tt'}(X_i; \beta_0) \quad (2.2)$$

for $i = 1, \dots, n$ and $1 \leq t, t' \leq T$. Here β_0 is a $p \times 1$ vector of unknown parameters. The functions $q_t(\cdot, \cdot)$ and $h_{tt'}(\cdot, \cdot)$ are fixed and known and may satisfy some constraints. For example, if Y_{it} is a dichotomous 0-1 variable, $q_t(X_i; \beta) = h_{tt}(X_i; \beta)$ and $h_{tt'}(X_i; \beta) \leq q_t(X_i; \beta)q_{t'}(X_i; \beta) + \{q_t(X_i; \beta)[1 - q_t(X_i; \beta)]q_{t'}(X_i; \beta)[1 - q_{t'}(X_i; \beta)]\}^{1/2}$. The restrictions (2.1) and (2.2) include models for mean and covariances. For example, if the parameters of the mean and covariance models vary independently and $E(Y_{it} | X_i) = \bar{q}_t(X_i; \alpha_0)$, $\text{cov}(Y_{it}, Y_{it'} | X_i) = c_{tt'}(X_i; \theta_0)$ then $\beta_0 = (\alpha_0^T; \theta_0^T)^T$, $q_t(X_i; \beta_0) = \bar{q}_t(X_i; \alpha_0)$ and $h_{tt'}(X_i; \beta_0) = c_{tt'}(X_i; \theta_0) + \bar{q}_t(X_i; \alpha_0)\bar{q}_{t'}(X_i; \alpha_0)$.

Our goal is to estimate β_0 when the full vector Y_i is not always observed because some subjects drop out of the study. We assume that the study is designed so that, in addition to Y_{it} and X_i , measurements are to be made on a $r \times 1$ vector of time dependent covariates V_{it} , $t = (0, \dots, T)$. We set $W_{it} = (V_{it}^T, Y_{it})^T$, $t = (1, \dots, T)$ and $W_{i0} = (X_i, V_{i0}, Y_{i0})$.

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 X_i is always observed and 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 , and that the missing data patterns are monotone, i.e. once a subject leaves the study, return is not possible, or, equivalently, $R_{it} = 0$ implies $R_{i(t+1)} = 0$. Define $\bar{W}_{it} = (W_{i0}, W_{it}, \dots, W_{i(t-1)})$. Here and throughout we shall use the convention that overbars are used to include past data recorded up to but not including the corresponding occasion. We shall assume that the missing data process satisfies

$$P(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{it}, Y_i) = P(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{it}) \tag{2.3}$$

and that

$$P(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{it}) > \sigma > 0, \quad t = 1, \dots, T \tag{2.4}$$

so that each subject i has a positive probability of remaining in the study. We shall suppose that the vectors $(\bar{W}_{i(T+1)}^T, R_{i1}, \dots, R_{iT})$, $i = 1, \dots, n$, are independent and identically distributed.

Under (2.3), censoring (non-response) is unrelated to current and future outcomes given the past \bar{W}_{it} . Assumption (2.3) holds in particular when the data are missing at random in the sense of Rubin (1976) since under monotonicity, missing at random is equivalent to

$$P(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{i(T+1)}) = P(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{it}). \tag{2.5}$$

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

$$\bar{\lambda}_{it} = \bar{\lambda}_{it}(\alpha_0). \tag{2.6}$$

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 .

For each subject we shall call $L_i = \bar{W}_{i(T+1)}$ the full data and

$$(R_{i1}, \dots, R_{iT}, L_{\text{obs}, i}^T) \tag{2.7}$$

the observed data where, $L_{\text{obs}, i}$ is the observed component of L_i , i.e. $L_{\text{obs}, i} = \bar{W}_{it}$ if $R_{it} = 0$ and $R_{i(t-1)} = 1$ and $L_{\text{obs}, i} = L_i$ if $R_{iT} = 1$. A semi-parametric model is characterized both by the available data and by restrictions on the joint distribution of the data. Our "full-data" semi-parametric model is defined by the restrictions (2.1) and (2.2) and the data L_i . Our "observed-data" semi-parametric model is defined by the restrictions (2.1)–(2.4) and (2.6) and data (2.7). The first goal of this paper is to propose a class of estimators of β_0 that are consistent and asymptotically normal under the "observed-data" semiparametric model. The second goal is to show that the asymptotic variance of the optimal estimator in our class attains the semi-parametric variance bound for all regular estimators of β_0 in the sense of Begun *et al.* (1983).

3. Estimation in the "observed-data" model

In this section we propose a class of estimators of β_0 . To help understand our estimators in the missing data problem it will be convenient to study first the problem of estimating β_0

when no outcomes are missing. If Y_{it} , $t = 1, \dots, T$, is not a dichotomous variable define $Z_{it} = (Y_{it}, Y_{it}^2, Y_{it}Y_{i(t-1)}, \dots, Y_{it}Y_{i1})^T$ and $(Z_i = Z_{i1}^T, \dots, Z_{iT}^T)^T$. Z_{it} and Z_i are column vectors of dimension $t+1$ and $T(T+3)/2$ respectively. Let $g_t(X_i; \beta_0) = (q_t(X_i; \beta_0), h_{t1}(X_i; \beta_0), h_{t(t-1)}(X_i; \beta_0), \dots, h_{t1}(X_i; \beta_0))^T$ be the vector of conditional means of Z_{it} given X_i . When Y_{it} is a dichotomous variable define $Z_{it} = (Y_{it}, Y_{it}Y_{i(t-1)}, \dots, Y_{it}Y_{i1})$ and $g_t(X_i; \beta) = (q_t(X_i; \beta), h_{t(t-1)}(X_i; \beta), \dots, h_{t1}(X_i; \beta))$. Define $\varepsilon_{it}(\beta) = Z_{it} - g_t(X_i; \beta)$ and $\varepsilon_i(\beta) = (\varepsilon_{i1}(\beta)^T, \dots, \varepsilon_{iT}(\beta)^T)^T$. Let $d(X_i, \beta)$ be a $p \times T(T+3)/2$ matrix of fixed functions of X_i and β ($d(X_i; \beta)$ is a $p \times T(T+1)/2$ matrix function if Y_{it} , $t = 1, \dots, T$, is dichotomous). Consider the estimating equations

$$U_{\text{full}}(\beta) = n^{-1/2} \sum_{i=1}^n d(X_i, \beta) \varepsilon_i(\beta) = 0. \quad (3.1)$$

Equation (3.1) includes the class of estimating equations for mean and covariance parameters proposed by Prentice (1988) for repeated binary outcomes and by Prentice & Zhao (1991) for multivariate continuous and discrete outcomes. These authors considered $d(X_i; \beta) = \{\partial g(X_i, \beta) / \partial \beta\} c(X_i)^{-1}$ where $c(X_i)$ is an arbitrary "working" covariance matrix of Z_i given X_i chosen by the investigator and $g(X_i; \beta) = (g_1(X_i; \beta)^T, \dots, g_T(X_i; \beta)^T)^T$.

When no outcomes are missing the estimating equation (3.1) has, under regularity conditions a solution $\hat{\beta}(d)$ that is consistent and asymptotically normal for estimating β_0 (Prentice & Zhao, 1991). The following lemma states that any regular and asymptotically linear estimator of β_0 is asymptotically equivalent to a solution of an equation in the class (3.1) for some choice of d . (The definitions of regularity and asymptotic linearity are given in the appendix.)

Lemma 1

Suppose $\hat{\beta}$ is a regular and asymptotically linear estimator of β_0 . Then there exists a function $d(X_i; \beta)$ such that $n^{1/2}\{\hat{\beta} - \hat{\beta}(d)\} = o_p(1)$.

Chamberlain (1987) showed that the asymptotic variance of $\hat{\beta}(d_{\text{full}})$ where $d_{\text{full}}(X_i; \beta) = \{\partial g(X_i, \beta) / \partial \beta\} \text{var}(\varepsilon_i | X_i)^{-1}$ and $\varepsilon_i = \varepsilon_i(\beta_0)$ attains the semi-parametric variance bound for the full data model. Since $\hat{\beta}(d_{\text{full}})$ does not use data on the time dependent covariates V_{it} , this implies that when no outcomes are missing, V_{it} does not asymptotically add information about β_0 .

Suppose now that some outcomes are missing and consider equation (3.1) restricted to the observed outcomes, i.e.

$$n^{-1/2} \sum d^*(X_i; \beta) \varepsilon_i^*(\beta) = 0. \quad (3.2)$$

Here, for each β , $\varepsilon_i^*(\beta)$ is equal to the vector of observed residuals and, given $d(X_i; \beta)$, $d^*(X_i; \beta)$ is the corresponding sub-matrix. Unfortunately, consistency of a solution of (3.2) typically requires the stronger condition that the data are missing completely at random (Rubin, 1976), i.e.

$$P(R_{it} = 1 | R_{i(t-1)} = 1, \bar{W}_{i(T+1)}) = P(R_{it} = 1 | R_{i(t-1)} = 1, X_i). \quad (3.3)$$

This is so since when (2.3) holds but (3.3) is false $\varepsilon_i^*(\beta_0)$ no longer has mean zero conditional on response because, as noted by Rubin (1976), subjects with $R_{it} = 1$ represent a biased sample.

In order to derive our estimators of β_0 let $\hat{\alpha}$ be the partial maximum likelihood estimator of α_0 in model (2.6), i.e. $\hat{\alpha}$ maximizes the partial likelihood,

$$L(\alpha) = \prod_i L_i(\alpha) = \prod_i \prod_t [\bar{\lambda}_{it}(\alpha)^{R_{it}} \{1 - \bar{\lambda}_{it}(\alpha)\}^{1 - R_{it}} R_{it}^{R_{it} - 1}] \tag{3.4}$$

or equivalently $\hat{\alpha}$ solves $\Sigma S_{xi}(\alpha) = 0$ where $S_{xi}(\alpha) = \{\partial \log L_i(\alpha) / \partial \alpha\} = \Sigma_{t=1}^T \{R_{it} - \bar{\lambda}_{it}(\alpha) R_{it}^{R_{it} - 1}\} \{\partial \log [1 - \bar{\lambda}_{it}(\alpha)] / \partial \alpha\}$ is the contribution to the score for α from the i th subject. Our estimators $\hat{\beta}(d, \phi)$ will be indexed by $d(X_i; \beta)$ a $p \times T(T + 3)/2$ matrix function of X_i for each β , $d(X_i; \beta)$ is a $p \times T(T + 1)/2$ matrix function if Y_{it} , $t = 1, \dots, T$, is dichotomous) and $\phi(\bar{W}_{iT}) = (\phi_1^T(\bar{W}_{i1}), \dots, \phi_T^T(\bar{W}_{iT}))^T$ where $\phi_j(\bar{W}_{ij})$ is a $p \times 1$ vector function of \bar{W}_{ij} . Specifically, $\hat{\beta}(d, \phi)$ solves

$$U(\beta, d, \phi, \hat{\alpha}) \equiv n^{-1/2} \sum_{i=1}^n U_i(\beta, d, \phi, \hat{\alpha}) = 0 \tag{3.5}$$

where $U_i(\beta, d, \phi, \alpha) = R_{iT} \bar{\pi}_{iT}(\alpha)^{-1} d(X_i; \beta) e_i(\beta) - A_i(\phi, \alpha)$ with

$$A_i(\phi, \alpha) = \sum \{R_{it} - \bar{\lambda}_{it}(\alpha) R_{it}^{R_{it} - 1}\} \bar{\pi}_{it}(\alpha)^{-1} \phi_t(\bar{W}_{it}) \quad \text{and} \quad \bar{\pi}_{it}(\alpha) = \prod_{s=1}^t \bar{\lambda}_{is}(\alpha).$$

The estimating equations (3.5) use data from drop-outs, i.e. subjects with $R_{iT} = 0$ as well as from subjects with full data. Drop-outs contribute to the estimating equations (3.5) through the estimator $\hat{\alpha}$ as well as through $A_i(\phi, \hat{\alpha})$. For example, if subject i leaves the study at time $t = 2$, then $A_i(\phi, \hat{\alpha})$ is equal to $\{1 - \bar{\lambda}_{i1}(\hat{\alpha})\} \bar{\lambda}_{i1}(\hat{\alpha})^{-1} \phi_1(\bar{W}_{i1}) - \bar{\lambda}_{i1}(\hat{\alpha})^{-1} \phi_2(\bar{W}_{i2})$. Note that $A_i(\phi, \alpha)$ is a function of the observed data since if a subject leaves the study at time t , the coefficients $R_{it} - \bar{\lambda}_{it}(\alpha) R_{it}^{R_{it} - 1}$ of the unobservables $\phi_{it}(\bar{W}_{it})$ are equal to zero for $t' > t$. Subjects with full data contribute additionally the term $\bar{\pi}_{iT}(\hat{\alpha})^{-1} d(X_i; \beta) e_i(\beta)$ which, when (2.5) holds, is equal to their contribution to the full-data estimating equation (3.1) weighted by the inverse of the estimate of their conditional probability of remaining in the study at the end of follow-up given $\bar{W}_{i(T+1)}$.

The following theorem gives the asymptotic properties of $\hat{\beta}(d, \phi)$ under the regularity conditions (R.1)–(R.9) of Appendix 1, which we shall henceforth assume to be true. In what follows it will be convenient to define $A^{\otimes 2} \equiv AA^T$ and $\text{resid}(A, B) \equiv A - E(AB^T) \text{var}(B)^{-1} B$, the residual from the population regression of A on B .

Theorem 1

Under (2.1)–(2.4) and (2.6), (i) with probability approaching 1 (w.p.a. 1) there exists a unique solution $\hat{\beta}(d, \phi)$ to (3.5); (ii) $n^{1/2}\{\hat{\beta}(d, \phi) - \beta_0\}$ is asymptotically normal with mean zero and covariance $I(d)^{-1} \Omega(d, \phi, \alpha_0) I(d)^{-1T}$ that can be consistently estimated by $\hat{I}(d)^{-1} \hat{\Omega}(d, \phi, \hat{\alpha}) \hat{I}(d)^{-1T}$ where, $I(d) = E[\partial \{d(X_i; \beta_0) e_i(\beta_0)\} / \partial \beta^T]$, $\Omega(d, \phi, \alpha_0) = E[\text{resid}\{U_i(\beta_0, d, \phi, \alpha_0), S_{x,i}\}^{\otimes 2}]$, $\hat{I}(d) = n^{-1} \sum R_{iT} \bar{\pi}_{iT}^{-1}(\hat{\alpha}) d(X_i; \hat{\beta}) \partial e_i(\hat{\beta}) / \partial \beta^T$, $\hat{\Omega}(d, \phi, \hat{\alpha}) \equiv n^{-1} \sum \text{resid}_i^{\otimes 2}$ and resid_i is the residual for subject i from the least squares regression of $U_i(\hat{\beta}, d, \phi, \hat{\alpha})$ on $S_{x,i}(\hat{\alpha})$.

Since the variance of the residuals from a least squares regression can never increase as the number of covariates increases, part (ii) of theorem 1 implies that given two nested correctly specified non-response models indexed by the parameters $\alpha^{(1)}$ and $\alpha^{(2)} = (\alpha^{(1)T}, \psi^T)^T$, $\Omega(d, \phi, \alpha^{(1)}) \geq \Omega(d, \phi, \alpha^{(2)})$ and hence the asymptotic variance of the estimator that uses the non-response model parametrized by $\alpha^{(2)}$ is no larger than that of the estimator that uses the non-response model parametrized by $\alpha^{(1)}$. Thus, increasing the dimension of the parameter vector α in the model (2.6) can never decrease and typically improves the efficiency with which we estimate β_0 .

Our next result states that the class of estimators $\hat{\beta}(d, \phi)$ contains, modulo $o_p(n^{-1/2})$, all regular and asymptotically linear estimators of β_0 under the ‘‘observed-data’’ semi-paramet-

ric model and that the asymptotic variance of a member of this class attains the semi-parametric variance bound.

Theorem 2

(i) Suppose that $\hat{\beta}$ is a regular and asymptotically linear estimator of β_0 in the "observed-data" semi-parametric model. Then there exist d and ϕ such that $n^{1/2}\{\hat{\beta} - \hat{\beta}(d, \phi)\} = o_p(1)$; (ii) There exist unique functions $d_{\text{eff}}(X_i; \beta)$ and $\phi_{\text{eff}}(\bar{W}_{iT}; \beta)$ such that $\Omega(d_{\text{eff}}, \phi_{\text{eff}}, \alpha_0)$ equals the semi-parametric variance bound in the "observed-data" semi-parametric model. Furthermore $I(d_{\text{eff}}) = \Omega(d_{\text{eff}}, \phi_{\text{eff}}, \alpha_0)$ so that the asymptotic variance of $\hat{\beta}(d_{\text{eff}}, \phi_{\text{eff}})$ attains the bound; (iii) $d_{\text{eff}}(X_i; \beta) = \partial g(X_i; \beta) / \partial \beta^T \text{var} \{R_{iT} \bar{\pi}_{iT}(\alpha_0)^{-1} \varepsilon_i - A_i(\mu, \alpha_0) \mid X_i\}^{-1}$ where $\mu(\bar{W}_{iT}) = (\mu_1(\bar{W}_{iT})^T, \dots, \mu_T(\bar{W}_{iT})^T)$ with $\mu_t(\bar{W}_{it}) = E(\varepsilon_i \mid \bar{W}_{it}, R_{it-1} = 1)$, and $\phi_{\text{eff}, t}(\bar{W}_{it}) = d_{\text{eff}}(X_i; \beta_0) \mu_t(\bar{W}_{it})$, $1 \leq t \leq T$.

Note that because the asymptotic variance of $\hat{\beta}(d_{\text{eff}}, \phi_{\text{eff}})$ is equal to the semi-parametric variance bound then, when d_{eff} and ϕ_{eff} are used in equation (3.5) instead of d and ϕ respectively, we can no longer improve the efficiency with which we estimate β_0 by increasing the degree of parametrization of the non-response model.

Suppose now that the data are missing completely at random, i.e. equation (3.3) holds, and let the semi-parametric model "random-observed-data" be defined as the "observed-data" model except that restriction (2.3) is replaced by (3.3). Because (3.3) implies (2.3) the solutions of equations (3.5) are also consistent and asymptotically normal for estimating β_0 under the "random-observed-data" model. In fact, the class of estimators that solves (3.5) still contains, modulo $o_p(1)$, all regular and asymptotically linear estimators of β_0 and furthermore, it has a member whose asymptotic variance achieves the semi-parametric variance bound in the "random-observed-data" model as the following theorem states.

Theorem 3

In the "random-observed-data" model the conclusions of theorem 2 remain true.

Theorem 3 implies that when we know that (2.3) holds, further knowledge that the non-response process satisfies the stronger condition that the data are missing completely at random does not asymptotically provide additional information about β_0 .

4. Adaptive estimation

The solution $\hat{\beta}(d_{\text{eff}}, \phi_{\text{eff}})$ is not directly useful for data analysis because $\mu_t(\bar{W}_{it})$ and $d_{\text{eff}}(X_i; \beta_0)$ 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 $\hat{\beta}$ we will estimate, in order, $\mu_t(\bar{W}_{it})$, $d_{\text{eff}}(X_i; \beta_0)$ and $\phi_{\text{eff}, t}(\bar{W}_{it})$. To estimate $\mu_t(\bar{W}_{it})$ we cannot simply regress the estimated residuals $\varepsilon_{ij}(\hat{\beta})$ on functions of \bar{W}_{it} among subjects observed at time j since the missing mechanism (2.3) does not imply that $E(\varepsilon_{ij} \mid \bar{W}_{it}, R_{it-1} = 1)$ equals $E(\varepsilon_{ij} \mid \bar{W}_{it}, R_{ij} = 1)$ for $j > t + 1$. Recall that $\varepsilon_{ij} \equiv \varepsilon_{ij}(\beta_0)$ where $\varepsilon_{ij}(\beta)$ is, by definition, the vector of residuals $(Y_{ij} - q_j(X_i; \beta), Y_{ij}^2 - h_{ij}(X_i; \beta), Y_{ij} Y_{i(j-1)} - h_{j(i-1)}(X_i; \beta), \dots, Y_{ij} Y_{i1} - h_{j1}(X_i; \beta))^T$. However, it is straightforward to prove that when (2.3) holds $E(\varepsilon_{ij} \mid \bar{W}_{it}, R_{it-1} = 1) = E(\bar{\pi}_{i(t-1)} \bar{\pi}_{ij}^{-1} \varepsilon_{ij} \mid R_{ij} = 1, \bar{W}_{it}) P(R_{ij} = 1 \mid R_{it-1} = 1, \bar{W}_{it})$ where $\bar{\pi}_{it} = \bar{\pi}_{it}(\alpha_0)$, $j \geq t$ and $1 \leq t \leq T$. Hence, we adopt the following two-stage estimation procedure.

At the first stage, we specify flexible regression models

$$P(R_{ij} = 1 \mid R_{it-1} = 1, \bar{W}_{it}) = l_j^{(i)}(X_j^{(i)}, \bar{W}_{it}), \quad (4.1)$$

$$E\{\bar{\pi}_{i(t-1)}\bar{\pi}_{ij}^{-1}\varepsilon_{ij} \mid R_{ij} = 1, \bar{W}_{it}\} = m_j^{(j)}(\tau_j^{(j)}, \bar{W}_{it}) \tag{4.2}$$

depending on finite dimensional parameters $\chi_j^{(j)}$ and $\tau_j^{(j)}$. Often the right-hand side of (4.1) will be chosen to be of logistic form. Given preliminary estimates $\hat{\beta}$ of β_0 and $\hat{\alpha}$ of α_0 , let $\hat{\chi}_j^{(j)}$ be the maximum likelihood estimator $\chi_j^{(j)}$ and let $\hat{\tau}_j^{(j)}$ be the (possibly non-linear) least squares estimator of $\tau_j^{(j)}$ 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 j th occasion. Then $\hat{\mu}_t(\bar{W}_{it}) = (\hat{\mu}_{t1}^T, \dots, \hat{\mu}_{tT}^T)^T$ where $\hat{\mu}_{tj} = l_j^{(j)}(\hat{\chi}_j^{(j)}, \bar{W}_{it})m_j^{(j)}(\hat{\tau}_j^{(j)}, \bar{W}_{it})$, if $t \leq j \leq T$ and $\hat{\mu}_{tj} = \varepsilon_{ij}(\hat{\beta})$ if $1 \leq j < t$; $A_i(\hat{\mu}, \hat{\alpha}) = \sum_{t=1}^T (R_{it} - \bar{\lambda}_{it}(\hat{\alpha})R_{i(t-1)})\bar{\pi}_{it}(\hat{\alpha})^{-1}\hat{\mu}_t(\bar{W}_{it})$. We then estimate the $p \times p$ matrix $\text{var}\{R_{iT}\bar{\pi}_{iT}(\alpha_0)^{-1}\varepsilon_i - A_i(\mu, \alpha_0) \mid X_i\}$. Given the multivariate (possibly non-linear) regression models

$$E(S_{ik}S_{ik'} \mid X_i) = l_{kk'}(\Psi, X_i) \tag{4.3}$$

for $1 \leq k \leq k' \leq T(T+3)/2$, where S_{ik} is the k th element of $R_{iT}\bar{\pi}_{iT}(\alpha_0)^{-1}\varepsilon_i - A_i(\mu, \alpha_0)$, estimate Ψ with $\hat{\Psi}$, the (possibly non-linear) multivariate least squares estimator of the regression of $\hat{S}_{ik}\hat{S}_{ik'}$ on X_i . Here \hat{S}_{ik} is the k th element of $R_{iT}\bar{\pi}_{iT}(\hat{\alpha})^{-1}\varepsilon_i(\hat{\beta}) - A_i(\hat{\mu}, \hat{\alpha})$. The estimate of $d_{\text{eff}}(X_i; \beta)$ is given by $\hat{d}_{\text{eff}}(X_i; \beta) = \{\partial g(X_i, \beta) / \partial \beta\} \text{var}\{R_{iT}\bar{\pi}_{iT}(\alpha_0)^{-1}\varepsilon_i - A_i(\mu, \alpha_0) \mid X_i\}^{-1}$ where $\text{var}\{R_{iT}\bar{\pi}_{iT}(\alpha_0)^{-1}\varepsilon_i - A_i(\mu, \alpha_0) \mid X_i\}$ is the $T(T+3)/2 \times T(T+3)/2$ symmetric matrix with (k, k') element $l_{kk'}(\hat{\Psi}, X_i)$. Notice that by restricting the functions $l_{kk'}(\cdot, X_i)$ to be such that the symmetric matrix with (k, k') element $l_{kk'}(\Psi, X_i)$ is positive definite for all values of Ψ we guarantee the positive definiteness of $\text{var}\{R_{iT}\bar{\pi}_{iT}(\alpha_0)^{-1}\varepsilon_i - A_i(\mu, \alpha_0) \mid X_i\}$. The estimate of $\phi_{\text{eff}, t}(\bar{W}_{it})$ is given by $\hat{\phi}_{\text{eff}, t}(\bar{W}_{it}) = \hat{d}_{\text{eff}}(X_i; \hat{\beta})\hat{\mu}_t(\bar{W}_{it})$.

Finally, at the second stage, the adaptive estimator of β_0 is $\hat{\beta}(\hat{d}_{\text{eff}}, \hat{\phi}_{\text{eff}})$, the solution of the equation (3.5) that uses \hat{d}_{eff} and $\hat{\phi}_{\text{eff}}$ instead of d and ϕ .

It is standard to show that when (4.1), (4.2) and (4.3) are correctly specified, $\hat{\beta}(\hat{d}_{\text{eff}}, \hat{\phi}_{\text{eff}})$ has the same asymptotic distribution as $\hat{\beta}(d_{\text{eff}}, \phi_{\text{eff}})$ (see for example Robins *et al.*, 1992). Furthermore, $\hat{\beta}(\hat{d}_{\text{eff}}, \hat{\phi}_{\text{eff}})$ will be asymptotically unbiased for β_0 even when (4.1), (4.2) or (4.3) are misspecified or even incompatible in the sense that there exists no joint distribution for the observable random variables compatible with (2.1), (2.2), (2.6) and the models (4.1)–(4.3). A consistent estimator of the asymptotic variance of $n^{1/2}\{\hat{\beta}(\hat{d}_{\text{eff}}, \hat{\phi}_{\text{eff}}) - \beta_0\}$ that is robust to misspecification of (4.1)–(4.3) is given by $\hat{I}(\hat{d}_{\text{eff}})^{-1}\hat{\Omega}(\hat{d}_{\text{eff}}, \hat{\phi}_{\text{eff}}, \hat{\alpha})\hat{I}(\hat{d}_{\text{eff}})^{-T}$ where the matrix functionals $\hat{I}(\cdot)$ and $\hat{\Omega}(\cdot, \cdot, \cdot)$ are as defined in theorem 1.

5. Final remarks

In this paper we have considered the efficient estimation of the parameters of models for mean and covariances of repeated outcomes under monotone missing at random or missing completely at random data. Our model assumes that the non-response probabilities are known up to a vector of unknown parameters. Using the results of Robins & Rotnitzky (1992) and Robins *et al.* (1994), it can be shown that the semi-parametric variance bound for estimating β_0 is unchanged if no restrictions are imposed on the non-response probabilities, and therefore that the non-response process does not asymptotically carry information about β_0 . When the non-response probabilities are not specified we can replace them, in the optimal estimating equation (3.5) that uses d_{eff} and ϕ_{eff} , by their kernel regression estimators. Under appropriate smoothness conditions and with the bandwidth appropriately chosen, this approach can be shown to result in semi-parametric efficient estimators of β_0 . However, unrealistically large samples are needed before the asymptotic distribution of the resulting estimator is a good approximation to its sampling distribution. Thus, in practice one requires models for the non-response probabilities.

The estimating equation (3.5) uses only the residuals $\varepsilon_i(\beta)$ of subjects with full data, i.e. with $R_{iT} = 1$. Thus, when $\bar{\pi}_{iT}(\beta_0)$ is small, the estimator $\hat{\beta}(d, \phi)$ may possibly have poor small

sample behaviour. As an alternative, one can replace in the estimating equation (3.5) the functions $\phi_t(\bar{W}_t)$ by functions $\phi_t^*(\bar{W}_t; \beta) = d(X)\bar{I}^{(t)}\varepsilon(\beta) + \phi_t(\bar{W}_t)$, where $\bar{I}^{(t)} = \text{diag}(\bar{e}_j^{(t)})$ satisfying $\bar{e}_j^{(t)} = 0$ if $j \geq t$ and $\bar{e}_j^{(t)} = 1$ if $j < t$. The resulting estimating equations use the residuals at all times t where an outcome is not missing. The asymptotic distribution of the resulting estimator coincides with that of the solution to equation (3.5) when $\phi_t(\bar{W}_t) = \phi_t^*(\bar{W}_t; \beta_0)$.

In this paper we have focused on the case of monotone missing data patterns. Our results can be extended to scenarios with non-monotone missing data patterns following the ideas in Robins *et al.* (1994, 1995). This extension is beyond the scope of this paper.

In this article we have considered efficient estimation of the parameters of models of means and covariances. Robins & Rotnitzky (1992), Robins *et al.* (1995), Robins & Rotnitzky (1995) and Heyting *et al.* (1992) consider the same problem when only models for the mean of the responses are known, i.e. when (2.1) is true but (2.2) is not specified. Robins & Rotnitzky (1992) extend these results by allowing censoring (dropout) to occur in continuous time.

Finally, caution is needed in using the methods of this paper. Although knowledge of the model for the second moments often helps in the estimation of mean parameters, the gain in efficiency of the semi-parametric efficient estimation (SEE) in models of means and covariances over that of the SEE in a semi-parametric model that imposes restrictions only on the first moments is accompanied by a lack of robustness to misspecification of the model for the second moments. A sensitivity analysis deserves further study.

Acknowledgements

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 US National Institutes of Health. Andrea Rotnitzky was additionally supported in part by a Mellon Foundation Faculty Development Award.

References

- Arminger, G. & Sobel, M. E. (1990). Pseudo-maximum likelihood estimation of mean and covariance structures with missing data. *J. Amer. Statist. Assoc.* **85**, 195–203.
- Begun, J. M., Hall, W. J., Huang, W. M. & Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11**, 432–452.
- Bickel, P., Klaassen, C. A. J., Ritov, Y. & Wellner, J. A. (1992). *Efficient and adaptive inference in semiparametric models*. Johns Hopkins University Press, Baltimore, MD.
- Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econom.* **34**, 305–324.
- DeGruttola, V., Lange, N. & Dafni, U. (1991). Modeling the progression of HIV infection. *J. Amer. Statist. Assoc.* **86**, 569–577.
- Gourieroux, C., Monfort, A. & Trognon, A. (1984). Pseudo maximum likelihood methods: theory. *Econometrica* **52**, 681–700.
- Heyting, A., Tolboom, J. T. B. M. & Essers, J. G. A. (1992). Statistical handling of drop-outs in longitudinal clinical trials. *Statistics in Medicine* **11**, 2043–2062.
- Manski, C. F. (1988). *Analog estimation methods in econometrics*. Chapman & Hall, New York.
- Newey, W. K. (1990). Semiparametric efficiency bounds. *J. Appl. Econom.* **5**, 99–135.
- Newey, W. K. & Powell, J. (1990). Efficient estimation of linear and type I censored regression models under conditional quantile restrictions. *Econometric Theory* **6**, 295–317.
- Newey, W. K. & McFadden, D. (1993). Estimation in large samples. *Handbook of econometrics*, Vol. 4 (eds D. McFadden & R. Engler). North Holland, Amsterdam.
- Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033–1048.

- Prentice, R. L. & Zhao, L. P. (1991). Estimating equations for parameters in mean and covariances of multivariate discrete and continuous responses. *Biometrics* **47**, 825–839.
- Robins, J. M. & Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology – methodological issues* (eds. N. Jewell, K. Dietz, & V. Farewell), pp. 297–331. Birkhäuser, Boston, MA.
- Robins, J. M. & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.*, **90**, 122–129.
- Robins, J. M., Mark, S. D. & Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48**, 479–495.
- Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when a regressor is not always observed. *J. Amer. Statist. Assoc.*, **89**, 846–866.
- Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes under the presence of missing data. *J. Amer. Statist. Assoc.* **90**, 106–121.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.
- Zhao, L. P. & Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642–648.

Received November 1993, in final form November 1994

Andrea Rotnitzky and James M. Robins, Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA.

Appendix

We first prove theorem 1. Let $H(\gamma)^T = (U_1(\beta, d)^T, S_x(\alpha)^T)$ where $U_1(\beta, d) = d(X; \beta)g(\beta)$, $\gamma^T = (\beta^T, \alpha^T)$ and $\gamma = \beta \times \alpha$ where β and α are the parameter spaces of β and α . We shall prove our propositions under the following nine regularity conditions.

- (R.1) γ lies in the interior of a compact set γ ;
- (R.2) (L_i, R_i) , $i = 1, \dots, n$, are independent and identically distributed;
- (R.3) $\lambda_{ii}(\alpha) > c > 0$ for all $\alpha \in \alpha$ for some c ;
- (R.4) $E[H(\gamma)] \neq 0$ if $\gamma \neq \gamma_0$;
- (R.5) $\text{var}[H(\gamma_0)]$ is finite and positive definite;
- (R.6) $E[\partial H(\gamma_0)/\partial \gamma']$ exists and is invertible;
- (R.7) $E[\sup_{\gamma \in N} \|H(\gamma)\|]$, $E[\sup_{\gamma \in N} \|\partial H(\gamma)/\partial \gamma'\|]$, $E[\sup_{\gamma \in N} \|H(\gamma)H(\gamma)'\|]$ are all finite, where $\|A\| \equiv \{\sum_{ij} A_{ij}^2\}^{1/2}$ for any matrix A with elements A_{ij} and N is a neighborhood of γ_0 ;
- (R.8) $f(L, R; \gamma)$ is a regular parametric model with score $S_\gamma(\gamma) = \partial \ln f(L, R; \gamma)/\partial \gamma$ where $f(L, R; \gamma)$ is a density that differs from the true density $f(L, R) = f(L, R; \gamma_0)$ only in that γ replaces γ_0 ;
- (R.9) For all γ^* in a neighborhood N of γ_0 , $E_{\gamma^*}[H(\gamma^*)]$ and $E_{\gamma^*}[\sup_{\gamma \in N} \|H(\gamma)'H(\gamma)\|]$ are bounded where E_{γ^*} refers to expectation w.r.t. the density $f(L, R; \gamma^*)$.

Proof of theorem 1. Let $U_i \equiv U_i(\beta_0, d, \phi, \alpha_0)$. First note that $E(U_i) = 0$ by (2.1)–(2.4) and (2.6). Let $H_i^*(\gamma)^T \equiv (U_i(\beta, d, \phi, \alpha)^T, S_{x,i}(\alpha)^T)$. Then, by $E(U_i) = 0$, regularity conditions (R.1)–(R.9) hold with $H^*(\gamma)$ instead of $H(\gamma)$. Th. (2.6) and (3.4) of Newey & McFadden (1993) or coroll. 1 in Ch. 8 of Manski (1988) imply that if $H^*(\gamma)$ satisfies (R.1)–(R.7) then w.p.a. 1 there exists a solution $\hat{\gamma}$ to $\sum H_i(\gamma) = 0$ such that $n^{1/2}(\hat{\gamma} - \gamma_0) = -E\{\partial H_i^*(\gamma_0)/\partial \gamma\}^{-1} n^{-1/2} \sum H_i^*(\gamma_0) + o_p(1)$ and part (i) of theorem 1 follows. By definition of $H_i^*(\gamma)$, $\hat{\gamma}^T = (\hat{\beta}(d, \phi)^T, \hat{\alpha}^T)$. Hence $n^{1/2}\{\hat{\beta}(d, \phi) - \beta_0\} = -E\{\partial U_i(\beta_0, d, \phi, \alpha_0)/\partial \beta\}^{-1} n^{-1/2} \sum [U_i - E\{\partial U_i(\beta_0, d, \phi, \alpha_0)/\partial \alpha^T\} E\{\partial S_{x,i}(\alpha_0)/\partial \alpha^T\}^{-1} S_{x,i}]$ where $S_{x,i} \equiv S_{x,i}(\alpha_0)$. Now by (2.3) $E\{\partial U_i(\beta_0, d, \phi, \alpha_0)/\partial \beta^T\} = E\{\partial U_{1,i}(\beta_0, d)/\partial \beta\} \equiv -I(d)$. Furthermore, under regularity conditions (R.6), (R.8) and (R.9), lem. (c.3) of Newey (1990) implies $-E\{\partial H_i^*(\gamma_0)/\partial \alpha^T\} = E\{H_i^*(\gamma_0)S_{x,i}(\alpha_0)^T\}$. Thus $E\{\partial U_i(\beta_0, d, \phi, \alpha_0)/\partial \alpha^T\} = -E\{U_i S_{x,i}^T\}$ and $E\{\partial S_{x,i}(\alpha_0)/\partial \alpha^T\} = E\{H_i^*(\gamma_0)S_{x,i}(\alpha_0)^T\}$. Thus $E\{\partial U_i(\beta_0, d, \phi, \alpha_0)/\partial \alpha^T\} = -E\{U_i S_{x,i}^T\}$ and $E\{\partial S_{x,i}(\alpha_0)/\partial \alpha^T\} = E\{H_i^*(\gamma_0)S_{x,i}(\alpha_0)^T\}$.

$\partial\alpha^T\} = -E(S_{x,t}S_{x,t}^T)$. Substituting these identities into the expression for $n^{1/2}\{\hat{\beta}(d, \phi) - \beta_0\}$ and using the Central Limit Theorem the asymptotic variance follows. The consistency of $\hat{I}(d)$ and $\hat{\Omega}(d, \phi, \hat{\alpha})$ is a direct consequence of the Law of Large Numbers.

We now prove lemma 1 and theorems 2 and 3. Our proofs use a general representation for the semi-parametric efficient score of an arbitrary semi-parametric model with monotone missing at random data given in th. 4.1 and 4.2 of Robins & Rotnitzky (1992) and prop. 8.1 and 8.2 of Robins *et al.* (1994), hereon denoted respectively by RR and RRZ. We start with a review of the theory of semi-parametric efficiency bounds borrowing heavily from the survey paper of Newey (1990) and the monograph of Bickel *et al.* (1992).

Suppose the data consists of n independent copies Z_i , $i = 1, \dots, n$, of a random variable Z . Let $L(\beta, \theta; Z_i)$ be the likelihood for a subject i in a semi-parametric 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 sub-model 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 semi-parametric 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 sub-model. The semi-parametric variance bound for β_0 is the supremum of the Cramer-Rao variance bounds for β_0 over all regular parametric sub-models. The asymptotic variance of any regular estimator of β_0 is no smaller than the semi-parametric variance bound.

Define the nuisance tangent space Λ to be the closed linear span of the set of all random vectors bS_η , where S_η is the subject-specific score for η evaluated at the truth in some regular parametric sub-model, 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. We shall consider Λ as a sub-set of the Hilbert space of $p \times 1$ random vectors H with inner product $E(H_1^T H_2)$ and $E(H^T H) < \infty$. 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 semi-parametric 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 semi-parametric model $L(\beta, \theta; Z)$, usually $S_\beta = \partial \ln L(\beta_0, \theta_0; Z)/\partial \beta$, and Λ^\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 B if $n^{1/2}(\hat{\beta} - \beta_0) = \sum_i B_i + o_p(1)$, $E(B) = 0$, and $\text{var}(B) < \infty$. In addition, Λ_0^\perp is the set of influence functions of regular asymptotic linear estimators of β_0 , where for any set \mathcal{F} of random variables, we define \mathcal{F}_0 to be the set with mean 0.

We now specialize the above general results to the semi-parametric model defined by (2.1)–(2.4) and data (2.7). 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 semi-parametric model characterized by (2.1) and (2.2), 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 respectively. Let Z be the observed data vector for a subject, i.e. $Z = (\bar{W}_t^T, R^T)^T$ if the subject is observed at occasion $t-1$ but not at occasion t , i.e. $R_{t-1} = 1$ and $R_t = 0$, and $Z = (\bar{W}_{T+1}^T, R^T)^T$ for a subject without missing data. Let $L(\beta, \theta, Z)$ be the likelihood for a single subject in the semi-parametric model characterized by (2.1)–(2.4) and (2.6) and data (2.7).

Proof of lemma 1. Let $K = k(Z^{(F)})$ with $E(K) = 0$. We first show that

$$\Pi(K | \Lambda^{(F)\perp}) = E(K\varepsilon^T | X)E(\varepsilon\varepsilon^T | X)^{-1}\varepsilon \tag{A.1}$$

Following analogous identical to those in lem. A.5 of Newey & Powell (1990) we have $\Lambda^{(F)} = \{B_1 + B_2 + B_3; B_1 = b_1(X), B_2 = b_2(\varepsilon, X), B_3 = b_3(\varepsilon, X, V)$ with $E(B_1) = 0, E(B_2 | X) = 0, E(\varepsilon B_2^T | X) = 0$ and $E(B_3 | \varepsilon, X) = 0\}$. Further, it is easy to show that $\Lambda_1^{(F)} = \{B_1\}, \Lambda_2^{(F)} = \{B_2\}$ and $\Lambda_3^{(F)} = \{B_3\}$ are mutually orthogonal. Now, $\Pi[K | \Lambda^{(F)\perp}] = K - \Pi[K | \Lambda^{(F)}]$. But $\Pi[K | \Lambda^{(F)}] = \Pi[K | \Lambda_1^{(F)}] + \Pi[K | \Lambda_2^{(F)}] + \Pi[K | \Lambda_3^{(F)}]$ by orthogonality of $\Lambda_1^{(F)}, \Lambda_2^{(F)}$ and $\Lambda_3^{(F)}$. From the definition of a projection, one can check that $\Pi[K | \Lambda_1^{(F)}] = K - E(K | X, \varepsilon), \Pi[K | \Lambda_2^{(F)}] = E(K | X, \varepsilon) - E(K | X) - E(K\varepsilon^T | X)E(\varepsilon\varepsilon^T | X)^{-1}\varepsilon,$ and $\Pi[K | \Lambda_3^{(F)}] = E(K | X) - E(K)$. So finally, $\Pi[K | \Lambda^{(F)}] = K - E(K) - E(K\varepsilon^T | X)E(\varepsilon\varepsilon^T | X)^{-1}\varepsilon$.

An immediate consequence is that

$$\Lambda_0^{(F)\perp} = \{d(X)\varepsilon: d(X) \text{ is a conformable matrix with } p \text{ rows}\} \tag{A.2}$$

and thus Lemma 1 follows.

Proof of theorem 2. First consider the semi-parametric model defined by (2.1)–(2.2), (2.4)–(2.6) and data (2.7). This model differs from the “observed-data” model only in that the data are missing at random, i.e. it imposes only the additional restriction $f(V_t | \bar{W}_t, \varepsilon) = f(V_t | \bar{W}_t, R_t = 1, \varepsilon)$. Since this condition is non-identifiable, i.e. it may be true whatever the distribution of the observables is, the allowable densities for the observable random variables in the semi-parametric models defined by restrictions (2.1)–(2.4) and (2.6), and (2.1), (2.2), (2.4)–(2.6) are the same. Thus, in particular the semi-parametric efficient score and semi-parametric variance bound for estimating β_0 are the same in both models. We henceforth assume that in addition to (2.1)–(2.4) and (2.6), (2.5) holds. It follows from prop. (8.1)(b) and (8.2)(c) of RRZ or th. (4.1)(d) and (4.2) of RR and (A.2) that $\Lambda_0^\perp = \{R_T \pi_T^{-1} U_1(\beta_0, d) - A(\phi, \alpha_0)\}$, proving part (i). Now S_{eff} is in Λ_0^\perp so $S_{\text{eff}} = R_T \pi_T^{-1} U_1(\beta_0, d_{\text{eff}}) - A(\phi_{\text{eff}}, \alpha_0) \equiv U(\beta_0, d_{\text{eff}}, \phi_{\text{eff}})$ for some d_{eff} and ϕ_{eff} . Hence $\text{var}(S_{\text{eff}}) = \Omega(d_{\text{eff}}, \phi_{\text{eff}})$. Further, $I(d_{\text{eff}}) = -E\{U_1(\beta_0, d_{\text{eff}})S_\beta^{(F)T}\} = E\{U(\beta_0, d_{\text{eff}}, \phi_{\text{eff}})S_{\text{eff}}^T\} = \text{var}(S_{\text{eff}})$ where the second equality is by prop. (8.1)(c.1) of RRZ or th. (4.1)(e) of RR. This concludes the proof of part (ii). Now, it follows from prop. (8.1)(e.1) and (8.2)(d) of RRZ or th. (4.1)(f) and (4.2) of RR that $\phi_{\text{eff}}(\bar{W}_t) = E\{d_{\text{eff}}(X; \beta_0)\varepsilon | \bar{W}_t, R_{t-1} = 1\}$ where $d_{\text{eff}}(X; \beta_0)$ is the unique conformable matrix $d(X)$ of p rows satisfying $S_{\text{eff}}^{(F)} = d(X)\varepsilon + \Pi[v\{d(X)\varepsilon\} | \Lambda^{(F)\perp}]$ with $v(B) = \Sigma_{t-1}^T (1 - \lambda_t) \bar{\pi}_t^{-1} \{B - E(B | \bar{W}_t, R_{t-1} = 1)\}$ for any random variable B . Now $S_{\text{eff}}^{(F)} = \{\partial g(X; \beta_0) / \partial \beta^T\} \text{var}(\varepsilon | X)^{-1} \varepsilon$ (Chamberlain, 1987) and using (A.1) from the proof of lemma 1, it follows that $d_{\text{eff}}(X; \beta_0)$ is the unique matrix function $d(X)$ satisfying

$$\{\partial g(X; \beta_0) / \partial \beta^T\} \text{var}(\varepsilon | X)^{-1} \varepsilon = d(X)\varepsilon + d(X)E\{v(\varepsilon)\varepsilon^T | X\} \text{var}(\varepsilon | X)^{-1} \varepsilon. \tag{A.3}$$

Upon multiplying both sides of (A.3) by ε and taking conditional expectations given X and solving for $d(X)$ we obtain $d(X) = \partial g(X; \beta_0) / \partial \beta^T [\text{var}(\varepsilon | X) + E\{\Sigma_{t-1}^T (1 - \lambda_t) \bar{\pi}_t^{-1} \text{var}(\varepsilon | \bar{W}_t, R_{t-1} = 1) | X\}]^{-1}$ and it is straightforward to verify that this coincides with the expression given for $d_{\text{eff}}(X; \beta_0)$ in the statement of the theorem. This concludes the proof of the theorem.

Proof of theorem 3. Theorem 3 is a direct consequence of th. (4.1)(f) of RR and prop. (8.1)(e.1) of RRZ which implies that, when the data are missing at random, the set of influence functions, the efficient score and the semi-parametric variance bound do not depend on a model for, or knowledge of, the non-response probabilities $\bar{\lambda}_t$.