Biometrika (1995), **82**, 4, *pp.* 805–20
Printed in Great Britain

Semiparametric regression estimation in the presence of dependent censoring

BY ANDREA ROTNITZKY

Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.

AND JAMES M. ROBINS

Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.

SUMMARY

We propose a semiparametric estimation procedure for estimating the regression of an outcome Y, measured at the end of a fixed follow-up period, on baseline explanatory variables X, measured prior to start of follow-up, in the presence of dependent censoring given X. The proposed estimators are consistent when the data are 'missing at random' but not 'missing completely at random' (Rubin, 1976), and do not require full specification of the complete data likelihood. Specifically, we assume that the probability of censoring at time t is independent of the outcome Y conditional on the recorded history up to t of a vector of time-dependent covariates that are correlated with Y. Our estimators can be used to adjust for dependent censoring and nonrandom noncompliance in randomised trials studying the effect of a treatment on the mean of a response variable of interest. Even with independent censoring, our methods allow the investigator to increase efficiency by exploiting the correlation of the outcome with a vector of time-dependent covariates.

Some key words: Follow-up; Missing data; Noncompliance; Randomised trial; Semiparametric efficiency bound; Surrogate marker.

1. INTRODUCTION

In both randomised and nonrandomised follow-up studies, it is often of interest to estimate the conditional mean of a response variable Y measured at the end of a fixed follow-up period as a function of explanatory variables X measured at the start of the study. Typically, in randomised studies, explanatory covariates will include a treatment arm indicator as well as other baseline characteristics such as age, race, sex and pre-treatment clinical studies.

The goal of this paper is to provide methods for estimating the regression of the response Y on the explanatory covariates X when Y is not completely observed because some subjects are censored, i.e. drop out of the study prior to end of follow-up.

When the probability of censoring depends only on the explanatory variables X, asymptotically unbiased estimators of the regression parameters β_0 can be obtained by a standard complete case analysis, i.e. the least squares regression of Y on X restricted to the uncensored subjects. In practice, the probability that a subject drops out of the study at time t will often depend on the recorded history $\bar{W}_t = (W_1, \ldots, W_{t-1})$ up to t of a vector of covariates W_j measured at time j that are correlated with Y. The W_j are often referred to as surrogates for Y (Prentice, 1989; Pepe, 1992). When censoring at t depends on $\bar{W}_t$, the least squares regression of Y on X restricted to uncensored subjects will be biased. In this paper we propose a class of estimators that remain asymptotically unbiased for β_0 provided that the probability of censoring at t given $(X, Y, \bar{W}_t)$ depends only on $(X, \bar{W}_t)$ and that one can correctly specify a parametric model, for example a logistic regression model, for the probability of censoring given $\bar{W}_t$ and X. These estimators are analogous to estimators first proposed by Robins & Rotnitzky (1992), who, however, allowed censoring to occur in continuous time. Heyting, Tolboom & Essers (1992) have also proposed the use of similar estimators for monotone missing data patterns.

When censoring does not depend on $\bar{W}_t$, the proposed estimator may be more efficient than the least squares estimator restricted to the uncensored subjects because our estimator exploits the correlation between the time-dependent covariates $\bar{W}_t$ and Y.

When the probability that a subject is censored depends on $\bar{W}_t$, likelihood methods can also provide valid inferences concerning the parameters β_0 of the regression of Y on X (Rubin, 1976) but these methods require correct specification of a parametric model for the joint distribution of Y, X and the time-dependent covariates W_t. These methods can be very nonrobust to misspecification of the likelihood and furthermore, with non-Gaussian data, they are typically computationally complex. In contrast, our methods are computationally simple and do not require specification of the joint distribution of Y, X and the W_t beyond the model of interest, that is the model for the conditional mean of Y given X. However, as mentioned above, our methods do require that we correctly specify a model for the probability of censoring at t given the $\bar{W}_t$ and X.

The paper is organised as follows. In § 2, we begin by studying the simpler problem of efficiently estimating the mean of an outcome variable Y when some Y's are missing and data on a time-independent surrogate variable W associated with Y are available. In § 3, we allow for time-varying surrogates W_t. In § 4, we extend our methods to estimate the parameters of the regression of Y on baseline covariates X. In § 5, we show that, when Y_i is Bernoulli, our results can be used to improve upon an estimator recently proposed by Pepe (1992). Further, we prove that there exists an 'estimator' in our class whose asymptotic variance attains the semiparametric variance bound for our model. However, this result cannot be directly utilised in data analysis, since the particular member of our class which attains the bound depends on the true but unknown probability law generating the data. Therefore, in § 6 we propose a modified estimator whose asymptotic variance should nearly attain the variance bound for the model. The final section contains some remarks and further considerations.

2. The estimation of a marginal mean

2·1. *A dichotomous surrogate*

Let Y_i $(i = 1, \ldots, n)$ be a univariate discrete outcome of interest measured on the ith subject and let W_i $(=0, 1)$ be a dichotomous surrogate variable. We assume that W_i is always observed but Y_i is missing for some subjects. We let R_i denote the indicator of response: $R_i = 1$ if Y_i is observed and $R_i = 0$ otherwise. This set-up is a special case of the longitudinal design studied in § 3. It is also discussed in the recent surrogate marker literature (Pepe, 1992). In that literature, Y_i is commonly an error-free measurement of an outcome of interest based on highly accurate but quite expensive diagnostic test, while W_i is a surrogate for Y_i based on a cheap but inaccurate test. This set-up is also familiar from the sample survey literature. In that setting, W_i is a background demographic variable such as sex obtained on all subjects and Y_i is an outcome of interest measured only on the subset of sampled subjects. However, as in the surrogate marker literature but in contrast to much of the sample survey literature, we shall regard (R_i, Y_i, W_i) $(i = 1, \ldots, n)$ as independent and identically distributed random vectors. The goal is to estimate $E(Y_i) = \beta_0$, the mean of Y_i, without making any assumptions about the joint distribution of Y_i and W_i. In particular we are not interested in the conditional mean of Y_i given W_i. When no Y_i's are missing, the sample average $\hat{\beta}_{\text{full}} = n^{-1} \sum Y_i$ is the maximum likelihood estimator of β_0 whether or not data on W_i are available.

We shall consider estimators of β_0 under the following assumptions about the missing data mechanism:

$$\text{pr}(R_i = 1 \mid Y_i, W_i) = \text{pr}(R_i = 1 \mid W_i), \tag{2·1}$$

$$\pi_i := \text{pr}(R_i = 1 \mid W_i) \geqslant \sigma \text{ with probability 1 for some } \sigma > 0, \tag{2·2}$$

$$\pi_i \text{ is known.} \tag{2·3}$$

Equation (2·1) implies the data are 'missing at random' in the sense of Rubin (1976). That is, within each level of W_i, the probability that Y_i is missing does not depend on Y_i. Equation (2·2) is needed to guarantee than an $n^{\frac{1}{2}}$-consistent estimator of β_0 exists (Robins, Rotnitzky & Zhao, 1994). For binary W_i, (2·2) is trivially satisfied whenever $\pi_i > 0$. Equation (2·3) will be true when, as in a sample survey or the surrogate marker example of Pepe (1992), selection is under the control of the investigator. If π_i depends on W_i, the average of the observed Y's,

$$\hat{\beta}_{\text{comp}} := \textstyle\sum_i Y_i / \sum_i R_i, \tag{2·4}$$

will in general be inconsistent for β_0, since the subjects with complete data constitute a biased sample (Rubin, 1976). Since, under (2·1)–(2·3), the data (Y_i, W_i) are multinomial with no restrictions on the probabilities, the maximum likelihood estimator $\hat{\beta}_{\text{MLE}}$ of $E(Y_i)$ is

$$\hat{E}(Y_i \mid W_i = 1)\hat{f}(W_i = 1) + \hat{E}(Y_i \mid W_i = 0)\hat{f}(W_i = 0), \tag{2·5}$$

where

$$\hat{E}(Y_i \mid W_i = w) = \textstyle\sum_i R_i Y_i I(W_i = w) / \sum_i R_i I(W_i = w)$$

and $\hat{f}(W_i = j)$ is $\sum I(W_i = j)/n$ $(j = 0{\cdot}1)$. It is easy to verify that $\hat{\beta}_{\text{MLE}}$ is also the solution to the inverse probability weighted estimating equations

$$0 = \sum_{i=1}^{n} \hat{\pi}_i^{-1} R_i (Y_i - \beta), \tag{2·6}$$

where

$$\hat{\pi}_i := \hat{\pi}(W_i) = \textstyle\sum_j R_j I(W_j = W_i) / \sum_j I(W_j = W_i)$$

is the maximum likelihood estimator of π_i when (2·3) is not imposed and thus π_i is unknown. Since $\hat{\beta}_{\text{MLE}}$ is not a function of π_i, it follows that $\hat{\beta}_{\text{MLE}}$ is also the maximum likelihood estimator of β_0 even when (2·3) is not imposed. When π_i is known, a consistent estimator of β_0 would be given by the solution $\hat{\beta}$ to the unbiased estimating equation

$$0 = \sum_{i=1}^{n} \pi_i^{-1} R_i (Y_i - \beta). \tag{2·7}$$

The estimating equation (2·7) is unbiased since, by (2·1) and (2·2),

$$E\{\pi_i^{-1}R_i(Y_i-\beta_0)\} = E\{E(R_i \mid W_i, Y_i)\pi_i^{-1}(Y_i-\beta_0)\} = E(Y_i-\beta_0)=0.$$

Inverse probability weighted estimating equations have been previously considered by Horvitz & Thompson (1952), Manski & Lerman (1977), Kalbfleisch & Lawless (1988), Flanders & Greenland (1991) and Zhao & Lipsitz (1992) among others. Weighting the observed data by $1/\pi_i$ is tantamount to allowing each subject with complete data to count for an additional $\pi_i^{-1}-1$ subjects with the same value of the covariate W_i for whom Y_i was not observed. Note that $\hat{\beta}$ has asymptotic variance at least as great as $\hat{\beta}_{\text{MLE}}$ since $\hat{\beta}_{\text{MLE}}$ is the maximum likelihood estimator. In Remark 1 below, we show that $\hat{\beta}$ is strictly less efficient than $\hat{\beta}_{\text{MLE}}$ unless $E(Y_i \mid W_i)=E(Y_i)$. Thus, one generates at least as precise, and usually more precise, an estimate of β_0 by using an estimate of $\text{pr}(R_i=1 \mid W_i)$ than by using the true population value, even were the latter known. However, even $\hat{\beta}_{\text{MLE}}$ can be highly variable and have poor finite sample performance if, for some subjects, π_i is nearly zero.

Suppose that π_i is a constant π, that is the data are 'missing completely at random' in the sense of Rubin (1976). Then $\hat{\beta}$ equals the complete case estimator $\hat{\beta}_{\text{comp}}$ in (2·4). Hence $\hat{\beta}_{\text{comp}}$ will be consistent for β_0 but will be strictly less efficient than the $\hat{\beta}_{\text{MLE}}$ unless $E(Y_i \mid W_i)=E(Y_i)$. However, when data on W_i are not recorded, $\hat{\beta}_{\text{comp}}$, which does not depend on W_i, becomes the maximum likelihood estimator of β_0. Thus, upon identifying the 'information' in an estimator with the inverse of its asymptotic variance,

$$[\text{var}^A\{m^{\frac{1}{2}}(\hat{\beta}_{\text{full}}-\beta_0)\}]^{-1}-[\text{var}^A\{m^{\frac{1}{2}}(\hat{\beta}_{\text{comp}}-\beta_0)\}]^{-1}$$

is the amount of the information about β_0 lost due to nonresponse in the absence of data on W_i. Here and throughout, var^A denotes asymptotic variance. Hence,

$$[\text{var}^A\{m^{\frac{1}{2}}(\hat{\beta}_{\text{MLE}}-\beta_0)\}]^{-1}-[\text{var}^A\{m^{\frac{1}{2}}(\hat{\beta}_{\text{comp}}-\beta_0)\}]^{-1}$$

is the amount of the lost information that can be recovered by measuring W_i. As an extreme but instructive example, suppose that Y_i is dichotomous and W_i is a perfect predictor of Y_i; that is $\text{var}(Y_i \mid W_i)=0$ with probability one. Then, by (2·5), $\hat{\beta}_{\text{MLE}}$ equals $\hat{\beta}_{\text{full}}$ and complete recovery of information is possible.

In summary, data on a surrogate W_i correlated with the missing outcome Y_i are always useful even when scientific interest is solely in the marginal distribution of Y_i. Even when Y_i is not discrete, $\hat{\beta}_{\text{MLE}}$ solving (2·6) remains the nonparametric maximum likelihood estimator of $E(Y_i)=\beta_0$, in the sense of Kiefer & Wolfowitz (1956). See also Gill (1989).

2·2. *Continuous surrogates*

Suppose now that W_i is vector-valued with multiple continuous components, with Y_i either continuous or discrete. Then completely nonparametric estimation of $E(Y_i \mid W_i=w)$ and $\pi_i=\text{pr}(R_i=1 \mid W_i)$ is not feasible due the 'curse of dimensionality' (Huber, 1985), and analogues of (2·5) or (2·6) require specification of models for either $E(Y_i \mid W_i=w)$ or π_i respectively to estimate $\beta_0=E(Y_i)$. In this paper we consider models for π_i. This choice (i) easily generalises to the regression set-up as discussed in § 4; (ii) is computationally straightforward; and (iii) is certain to be based on correctly specified models when, as in the sample survey setting, or surrogate marker examples of Pepe (1992), the selection probabilities of π_i are under the control of the investigator and thus are known. Specifically, we shall suppose we have a correctly specified parametric model $\pi_i(\alpha)$ for π_i. That is

$$\pi_i=\pi_i(\alpha_0), \tag{2·8}$$

where α_0 is an unknown parameter vector, $\pi_i(\alpha) := \pi(W_i; \alpha)$ is a known smooth function taking values in $(0, 1]$, and α and β are variation-independent. Typically $\pi_i(\alpha)$ is taken to be a logistic function; i.e.

$$\operatorname{logit} \pi_i(\alpha) = \alpha^{\mathrm{T}} H_i, \tag{2·9}$$

with $H_i = h(W_i)$, a vector-valued function of W_i. Let $\hat{\alpha}$ be the maximum likelihood estimator of α, satisfying

$$0 = \sum_{i=1}^{n} S_{\alpha,i}(\hat{\alpha}), \tag{2·10}$$

with

$$S_{\alpha,i}(\alpha) = \partial \log [\pi_i(\alpha)^{R_i}\{1 - \pi_i(\alpha)\}^{1-R_i}]/\partial\alpha = \{R_i - \pi_i(\alpha)\}\, \partial \operatorname{logit} \pi_i(\alpha)/\partial\alpha.$$

Now define $U_i(\beta, \alpha) = \{\pi_i(\alpha)\}^{-1} R_i \varepsilon_i(\beta)$ with $\varepsilon_i(\beta) = Y_i - \beta$, and let $\tilde{\beta}(\alpha)$ be a solution to $\sum U_i(\beta, \alpha) = 0$. Note that $\hat{\beta}$ solving (2·7) is $\tilde{\beta}(\alpha_0)$. Define $A^{[2]} = AA'$. We then have the following two propositions. A proof of the first is sketched in the Appendix and the second at the end of this section. Throughout, we assume that the regularity conditions (i)–(vii) in the Appendix are satisfied.

PROPOSITION 1. *If* (2·1), (2·2) *and* (2·8) *are true, then* (i) *with probability approaching* 1, *$\tilde{\beta}(\alpha_0)$ and $\tilde{\beta}(\hat{\alpha})$ exist and are unique;* (ii) *$n^{\frac{1}{2}}\{\tilde{\beta}(\alpha_0) - \beta_0\}$ and $n^{\frac{1}{2}}\{\tilde{\beta}(\hat{\alpha}) - \beta_0\}$ are asymptotically normal with mean* 0 *and variance $\tau^{-1}V\tau^{-1\prime}$ and $\tau^{-1}C\tau^{-1\prime}$ respectively, that can be consistently estimated by $\tilde{\tau}^{-1}\tilde{V}\tilde{\tau}^{-1\prime}$ and $\tilde{\tau}^{-1}\tilde{C}\tilde{\tau}^{-1\prime}$, where*

$$\tau = E\{\partial U_i(\beta_0, \alpha_0)/\partial\beta'\}, \quad V = \operatorname{var}(U_i), \quad U_i = U_i(\beta_0, \alpha_0),$$
$$C = \operatorname{resid}(U_i, S_{\alpha,i}), \quad S_{\alpha,i} = S_{\alpha,i}(\alpha_0),$$

with $\operatorname{resid}(A, B) = A - E(AB')\{E(BB')\}^{-1}B$ *the residual from the population least squares regression of A on B,*

$$\tilde{\tau} = n^{-1}\sum \partial U_i(\tilde{\beta}, \hat{\alpha})/\partial\beta', \quad \tilde{\beta} = \tilde{\beta}(\hat{\alpha}), \quad \tilde{V}_i = n^{-1}\sum \{U_i(\tilde{\beta}, \hat{\alpha})\}^{[2]},$$
$$\tilde{C} = n^{-1}\sum \{\operatorname{Resid}(U_i, S_{\alpha,i})\}^{[2]},$$

with $\operatorname{Resid}(U_i, S_{\alpha,i})$ *the residual for subject i from the least squares regression of the $U_i(\tilde{\beta}, \hat{\alpha})$ on the $S_{\alpha,i}(\hat{\alpha})$ $(i = 1, \ldots, n)$.*

It is easy to calculate that τ in Proposition 1 equals 1 and thus need not actually be estimated. However, the estimator $\tilde{\tau}$ will be needed later when we extend the estimation method to the regression setting.

PROPOSITION 2. *If* (2·1), (2·2) *and* (2·8) *hold, then:*
(i) $\operatorname{var}^A [n^{\frac{1}{2}}\{\tilde{\beta}(\hat{\alpha}) - \beta_0\}] \leqslant \operatorname{var}^A [n^{\frac{1}{2}}\{\tilde{\beta}(\alpha_0) - \beta_0\}]$;
(ii) $\operatorname{var}^A [n^{\frac{1}{2}}\{\tilde{\beta}(\hat{\alpha}) - \beta_0\}] \geqslant \tau^{-1} \operatorname{var}(U_i - S_i)\tau^{-1\prime}$, *where* $S_i = (R_i - \pi_i)\{E(\varepsilon_i | W_i)/\pi_i\}$ *with strict inequality unless*

$$b\, \partial \operatorname{logit} \pi_i(\alpha_0)/\partial\alpha = E(\varepsilon_i | W_i)/\pi_i \tag{2·11}$$

for some constant matrix b;

(iii) $\tau^{-1}\operatorname{var}(U_i - S_i)\tau^{-1\prime}$ *is the semiparametric variance bound for* β_0 *in the sense of Begun et al.* (1983).

Since (2·1), (2·2) and (2·8) impose no restrictions on the joint distribution of Y and W, part (iii) means that there is no regular estimator with asymptotic variance less than $\tau^{-1}\operatorname{var}(U_i - S_i)\tau^{-1\prime}$ that is guaranteed to be asymptotically normal and unbiased for β_0 whatever be the distribution of (Y, W) (Begun et al., 1983). Further, it follows from Proposition 6 below that the bound is unchanged if we do not impose the restriction (2·8) and thus π_i is completely unknown, or if we impose the additional restriction (2·3) that π_i is completely known.

Remark 1. When W_i is dichotomous, $\hat{\beta}_{\mathrm{MLE}}$ is $\tilde{\beta}(\hat{\alpha})$ with $\pi_i(\alpha)$ given by (2·9) with $H_i = (1, W_i)'$; further, (2·11) is true since any function of W_i, including $E(\varepsilon_i \mid W_i)/\pi_i$, can be written as a linear combination of $(1, W_i)' = \partial \operatorname{logit} \pi_i(\alpha_0)/\partial\alpha$. Hence $\tilde{\beta}(\alpha_0)$ and $\tilde{\beta}(\hat{\alpha})$ will have the same limiting distribution if and only if S_i is zero. But S_i is zero if and only if

$$E(U_i \mid W_i) \equiv E(Y_i - \beta_0 \mid W_i) \equiv E(Y_i \mid W_i) - E(Y_i) = 0.$$

Suppose now that, given a linear logistic model (2·9) for π_i, we add additional terms such as powers of the components of H_i and their interactions. This will increase the number of free parameters. As noted by Robins, Mark & Newey (1992) in a similar set-up, as we increase the number of parameters we may derive two benefits. First, we may reduce any large sample bias in $\tilde{\beta}(\hat{\alpha})$ due to misspecification of the model for π_i. Secondly, as stated in the following proposition, when the original model (2·9) is correctly specified, the efficiency with which we estimate β_0 will never decrease and will often increase. Now let $\alpha^{(j)}$ $(j = 1, \ldots, J)$ denote the parameter vector in the jth of a finite sequence of J correctly specified nested logistic models $\pi_i(\alpha^{(j)})$ for π_i indexed by increasing dimension $v^{(j)}$ of $\alpha^{(j)}$, where by convention we distinguish different models for π_i by their parameter vectors. By nesting the logistic models, we mean that the covariate vector $H_i^{(j)}$ of the jth model is a subvector of $H_i^{(j+1)}$. Correct specification implies that all but the first $v^{(1)}$ components of each $v^{(j)}$-vector of true values $\alpha_0^{(j)}$ are zero. In the Appendix, we prove the following.

PROPOSITION 3. *The asymptotic variance* $\operatorname{var}^A [n^{\frac{1}{2}}\{\tilde{\beta}(\hat{\alpha}^{(j)}) - \beta_0\}]$ *is nonincreasing with* j.

Thus it would be advantageous to use a richly parametrised model $\pi_i(\alpha)$ because this may help to guard against specification bias and may increase the efficiency of $\tilde{\beta}(\hat{\alpha})$. However, this must be tempered by two facts. First, the asymptotic efficiency of $\tilde{\beta}(\hat{\alpha})$ cannot exceed the bound in part (ii) of Proposition 2. Secondly, our proof that $\tilde{\beta}(\hat{\alpha})$ is asymptotically normal and unbiased for β_0 assumes $\pi_i(\hat{\alpha})$ to be $n^{\frac{1}{2}}$-consistent for π_i, which limits the number of free parameters in $\pi_i(\alpha)$. If model (2·9) for π_i is misspecified, the estimator $\tilde{\beta}(\hat{\alpha})$ may be more biased than the complete case estimator. One might overcome some of the sensitivity to model misspecification by allowing the dimensions v of the covariate vector H_i and parameter vector α to increase with the sample size n. Theorem (7.1) of Newey (1995) provides smoothness and regularity conditions guaranteeing that, if v increases with sample size, $\tilde{\beta}(\hat{\alpha})$ can achieve the efficiency bound given in part (iii) of Proposition 2 without sacrificing the regularity and asymptotic normality of the estimator. Specifically, if W_i has d continuous components, $\pi(W_i)$ has s continuous derivatives in these components of W_i, and the H_i are a power series basis, it is sufficient to choose v such that $v^6/n \to 0$ and $nv^{4-2s/d} \to 0$. Of course, as discussed by Huber (1985), if the dimension of W_i is large, this result may have limited application with the sample sizes found in practice due to the 'curse of dimensionality' and some model for π_i will be required.

Proof of Proposition 2. Part (i) follows from Proposition 1 and the fact that the residual variance from a regression is less than or equal to the variance of the independent variable. Hence $C \leqslant V$. Part (ii) follows from the fact that var $(U_i - S_i)$ minimises var $(U_i - G_i)$ over all $G_i \in G^* := \{(R_i - \pi_i)g(W_i)\}$ with $g(W_i)$ an arbitrary real-valued function, since

(a) $$\operatorname{var}(U_i - G_i) = \operatorname{var}\{(U_i - S_i) + (S_i - G_i)\}$$
$$= \operatorname{var}(U_i - S_i) + \operatorname{var}(S_i - G_i) + 2\operatorname{cov}\{(U_i - S_i), (S_i - G_i)\},$$

and

(b) $\operatorname{cov}\{(U_i - S_i), (S_i - G_i)\} = 0$ for all $G_i \in G^*$.

But, with $b^* := E(U_i S'_{\alpha,i})\{E(S_{\alpha,i} S'_{\alpha,i})\}^{-1}$, C equals var $(U_i - b^* S_{\alpha,i})$ and $b^* S_{\alpha,i} \in G$. Hence, $C \geqslant \operatorname{var}(U_i - S_i)$. If b satisfies (2·11), then $bS_{\alpha,i} = S_i$ and $C = \operatorname{var}(U_i - b^* S_{\alpha,i}) \leqslant \operatorname{var}(U_i - bS_{\alpha,i})$. Thus, $C = \operatorname{var}(U_i - S_i)$. Part (iii) is a corollary of Proposition 6. □

Results in this section are related to and generalise results of Rosenbaum (1987).

3. A longitudinal study

Suppose now that the design of the study calls for surrogate variables W_{it} to be measured at fixed times $t = 1, 2, \ldots, T-1$, and for Y_i, the outcome variable of interest, to be measured at time T, the end of follow-up. Let $\bar{W}_{it}$ record the ith subject's W-history up to but not including the tth occasion so that $\bar{W}_{it} = (W_{i0}, W_{i1}, \ldots, W_{i(t-1)})'$, where, for notational convenience, $W_{i0} = 0$. The goal of the study is to estimate the mean of Y_i without making any assumptions about the joint distribution of $(Y_i, \bar{W}_{iT})$. We let $R_{it} = 1$ if subject i is observed at the tth occasion $1 \leqslant t \leqslant T$ and $R_{it} = 0$ otherwise. We assume that missing patterns are monotone, i.e., once a cycle is missed, return to the study is not possible, or equivalently $R_{it} = 0$ implies $R_{i(t+1)} = 0$ and, by convention, $R_{i0} = 1$. The following assumptions concerning the nonresponse mechanisms generalise (2·1) and (2·2). For $1 \leqslant t \leqslant T$, among those at risk to be observed at time t, that is $R_{i(t-1)} = 1$, missingness is unrelated to the response Y_i conditional on the past, i.e.

$$\operatorname{pr}(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{it}, Y_i) = \operatorname{pr}(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{it}), \tag{3·1}$$

and, with probability one,

$$\bar{\lambda}_{it} := \operatorname{pr}(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{it}) \geqslant \sigma > 0. \tag{3·2}$$

Here and throughout we use overbars to indicate functions of the past history $\bar{W}_{it}$.

To generalise (2·8), we specify parametric models for the $\bar{\lambda}_{it}$:

$$\bar{\lambda}_{it} = \bar{\lambda}_{it}(\alpha_0) \quad (t = 1, \ldots, T), \tag{3·3}$$

where α_0 is an unknown parameter vector and $\bar{\lambda}_{it}(\alpha) = \lambda_{it}(\bar{W}_{it}; \alpha)$ is a known function taking values in (0, 1] with two continuous derivatives with respect to α. Typically $\bar{\lambda}_{it}(\alpha)$ is logistic, i.e.

$$\operatorname{logit} \bar{\lambda}_{it}(\alpha) = \alpha' \bar{H}_{it}, \tag{3·4}$$

where $\bar{H}_{it} = h(\bar{W}_{it})$.

Noting that, for a subject i lost to follow-up at time $t < T$, we have $R_{i(t-1)} - R_{it} = 1$, the model defined by the restriction (3·1)–(3·3) can be viewed as a semiparametric model for

β with likelihood

$$\prod_{i=1}^{n} \mathscr{L}_i^{\text{obs}}(\beta, \theta_1, \theta_2)\mathscr{L}_i^{\text{mis}}(\alpha), \tag{3·5}$$

where

$$\mathscr{L}_i^{\text{obs}}(\beta, \theta_1, \theta_2) = \{\mathscr{L}_i^{\text{full}}(\beta, \theta_1, \theta_2)\}^{R_{iT}} \times \prod_{t=1}^{T} \left\{ \int \mathscr{L}_i^{\text{full}}(\beta, \theta_1, \theta_2)\, dW_{it}, \ldots, dW_{iT-1}\, dY_i \right\}^{R_{i(t-1)} - R_{it}},$$

and

$$\mathscr{L}_i^{\text{full}}(\beta, \theta_1, \theta_2) = f_\varepsilon\{\varepsilon(\beta); \theta_1\} \prod_{t=1}^{T-1} f\{W_{it} \mid \bar{W}_{it}, \varepsilon(\beta), R_{it} = 1; \theta_{2t}\}$$

is the likelihood for a subject with complete data, and

$$\mathscr{L}_i^{\text{mis}}(\alpha) = \prod_{t=1}^{T} [\bar{\lambda}_{it}(\alpha)^{R_{it}}\{1 - \bar{\lambda}_{it}(\alpha)\}^{1-R_{it}}]^{R_{i(t-1)}}.$$

The parameters α and β are variation independent; θ_1 takes values in the set of all densities for ε_i with mean 0; θ_2 equals $(\theta'_{21}, \ldots, \theta'_{2T})'$ with θ_{2t} taking values in the set of all conditional densities for W_{it} given $(\bar{W}_{it}, \varepsilon_i, R_{i(t-1)} = 1)$. The parameter θ_1 is 'infinite dimensional' if Y_i is continuous; θ_2 is 'infinite-dimensional' if either Y_i or any component of $\bar{W}_{it}$ is continuous.

Let $\hat{\alpha}$ maximise the partial likelihood $\mathscr{L}_i^{\text{mis}}(\alpha)$. Then $\hat{\alpha}$ still solves (2·10) except now

$$S_{\alpha,i}(\alpha) = \frac{\partial \log\{\mathscr{L}_i^{\text{mis}}(\alpha)\}}{\partial\alpha} = \prod_{t=1}^{T} \{R_{it} - \bar{\lambda}_{it}(\alpha)R_{i(t-1)}\} \frac{\partial \operatorname{logit} \bar{\lambda}_{it}(\alpha)}{\partial\alpha}.$$

We define

$$\bar{\pi}_{it}(\alpha) = \prod_{j=1}^{t} \bar{\lambda}_{ij}(\alpha), \quad \bar{\pi}_{it} = \prod_{j=1}^{t} \bar{\lambda}_{ij}.$$

In the Appendix, we prove the following.

PROPOSITION 4. *If* (3·1)–(3·3) *hold then the conclusions of Propositions* 1–3 *remain true, with* $U_i(\beta, \alpha) := \{\bar{\pi}_{iT}(\alpha)\}^{-1} R_{iT}\varepsilon_i(\beta)$,

$$S_i := \sum_{t-1}^{T} \{R_{it} - \bar{\lambda}_{it} R_{i(t-1)}\}\bar{Q}_{it},$$

$$\bar{Q}_{it} := q_t(\bar{W}_{it}) := E\{\varepsilon_i \mid \bar{W}_{it}, R_{i(t-1)} = 1\}/\bar{\pi}_{it},$$

(2·11) *replaced by* $b\, \partial \operatorname{logit} \bar{\lambda}_{it}(\alpha_0)/\partial\alpha = \bar{Q}_{it}$ $(t = 1, \ldots, T)$ *and* j *now indexing nested logistic models* (3·4) *for* $\bar{\lambda}_{it}$.

Part (iii) of Proposition 2 now implies that there is no regular estimator of β_0 with asymptotic variance less than τ^{-1} var $(U_i - S_i)\tau^{-1\prime}$ that is guaranteed to be asymptotically normal and unbiased whatever be the nuisance parameters θ_1, θ_2 and α in (3·5). In § 6, we propose an adaptive estimator whose asymptotic variance may nearly attain the bound.

In analogy with § 2·1, suppose that, for each t, W_{it} is dichotomous. Then one can show that $\tilde{\beta}(\hat{\alpha})$ is the nonparametric maximum likelihood estimator of β_0 based on likelihood (3·5) when $\hat{\alpha}$ is obtained from the fit of the saturated logistic model logit $\bar{\lambda}_{it}(\alpha) = \alpha'_t \bar{H}_{it}$, where $\alpha = (\alpha_1', \ldots, \alpha_T')'$ and $\bar{H}_{it}$ is the 2^{t-1}-vector consisting of indicators for the possible realisations of $\bar{W}_{it}$. In further analogy with § 2, suppose missingness is completely at random in the sense that, in addition to (3·1), it is also true that

$$\mathrm{pr}\,(R_{it} = 1 \mid R_{i(t-1)} = 1, \bar{W}_{it}) = \mathrm{pr}\,(R_{it} = 1 \mid R_{i(t-1)} = 1). \tag{3·6}$$

Then $\hat{\beta}_{\mathrm{comp}} := \sum Y_i R_{iT} / \sum R_{iT}$ is $\tilde{\beta}(\alpha_0)$ and, thus, by Proposition 4, will in general be less efficient than $\tilde{\beta}(\hat{\alpha})$.

The results of this section can be used to provide estimates of the treatment effect in randomised trials in the presence of dependent censoring and nonrandom noncompliance. For purposes of illustration, consider a typical eight week randomised placebo controlled trial of a new antidepressive medication, say drug A. The contrast of primary interest in such a trial is the difference in the treatment-arm-specific means of a clinical depression score Y_i obtained at the end of the follow-up. Subjects are asked to return to the clinic each week to have the severity of their depression scored and to have various laboratory tests. We shall suppose that, during the course of the trial, a new antidepressive drug B receives intense media attention. Within each treatment arm, subjects with persistent depression during the follow-up period are more likely than others to initiate therapy with drug B and/or to fail to return to the clinic for further evaluation. Furthermore, suppose the trial cannot be successfully blinded, because most subjects suffer nausea and vertigo for a week or so upon initiating treatment with drug A, and that the proportion of subjects failing to comply with their assigned protocol by initiating therapy with drug B and/or by dropping out of the study is considerably greater in the placebo arm than in the treatment arm. To obtain some useful information regarding the benefits of treatment A as compared to placebo, it is agreed to regard a subject as censored at time t if (a) therapy with B has been initiated, or if (b) a clinic visit, and thus a urine test for drug B, has been missed at or before t. Suppose further that, among subjects in a given treatment arm with identical clinical and laboratory disease histories up to the $(t-1)$th visit, the decision to initiate therapy with drug B and the decision not to return to clinic in week t are unrelated to the outcome of interest, i.e. to the value of Y_i that would have been obtained at the 8th visit in the absence of therapy B or drop-out. This assumption, while never precisely true, is always less restrictive than the assumption that drop-out was completely at random. See Robins (1992, § 2) for further discussion of these issues.

In this setting, if we let W_{it} record the ith subject's depression score and laboratory values obtained at week t $(t \leqslant 7)$, then, under the assumption of the preceding paragraph, (3·1) will be true within each treatment arm separately but (3·6) will be false. Separately in each treatment arm, we can estimate, by $\tilde{\beta}(\hat{\alpha})$, the mean of the value of Y_i that would have been obtained at the 8th visit in the absence of therapy B or drop-out. An estimate of the treatment-arm effect is then obtained as the difference in the treatment-arm-specific estimates $\tilde{\beta}(\hat{\alpha})$. Even if, within each treatment arm, censoring is independent of Y_i in the sense that (3·6) holds, estimates of the treatment arm effect based on differences in $\tilde{\beta}(\hat{\alpha})$ are more efficient than those based on differences in the complete-case estimators $\hat{\beta}_{\mathrm{comp}}$.

4. Extensions to regression models

Suppose that in the longitudinal study of § 3 we additionally record, for subject i $(i = 1, \ldots, n)$, data on X_i, a vector of explanatory covariates measured at baseline prior to the start of follow-up. Suppose that the conditional mean of Y_i given X_i follows the regression model

$$E(Y_i | X_i) = g(X_i, \beta_0) \tag{4·1}$$

for $i = 1, \ldots, n$, where β_0 is a $p \times 1$ vector of unknown parameters and $g(x, \beta)$ is a known function with a continuous derivative with respect to β. Our goal is to estimate β_0 when Y_i is not always observed because some subjects drop out of the study prior to time t. We can use model (4·1) to describe the trial of § 3 without having to analyse the two treatment arms separately. If we let X_i^* be a dichotomous treatment arm indicator and set

$$g(X_i, \beta) = \beta_1 + \beta_2 X_i^*, \quad X_i = (1, X_i^*)',$$

then β_2 is the treatment-effect estimated in § 3. To look further for interactions between treatment and various pretreatment variables such as age, we can set

$$g(X_i, \beta) = \beta_1 + \beta_2 X_i^* + \beta_3 \, \text{age}_i + \beta_4 X_i^* \, \text{age}_i,$$

so $X_i = (X_i^*, \text{age}_i, X_i^* \, \text{age}_i)'$. The coefficient β_4 represents a treatment-arm age interaction. It is important to note that we do not assume $E(Y_i | X_i, \bar{W}_{it}) = E(Y_i | X_i)$ holds for any t. Rather we suppose it is the conditional mean of Y_i given X_i alone that is of substantive interest as would be the case in a randomised clinical trial.

Now redefine $\varepsilon_i(\beta) = Y_i - g(X_i, \beta)$ and set $W_{i0} := X_i$. The model characterised by the restrictions (3·1)–(3·3) and (4·1) is a semiparametric model for β_0 with likelihood function given by (3·5) except that now $f_\varepsilon\{\varepsilon_i(\beta); \theta_1\}$ is replaced by

$$f(X_i; \theta_{11}) f_\varepsilon\{\varepsilon_i(\beta) | X_i; \theta_{12}\}, \quad \theta_1 = (\theta_{11}, \theta_{12}).$$

The parameter θ_{11} takes values in the set of all densities for X_i, and θ_{12} takes values in the set of conditional densities for ε_i given X_i with mean zero. Let $D_i := d(X_i)$ be a $p \times 1$ vector of known functions of X_i. As noted by Rubin (1976), the estimating function $\sum_i D_i \varepsilon_i(\beta)$ may not have a consistent root when we do not have

$$f(Y_i | X_i, R_{iT} = 1) \equiv f(Y_i | X_i), \tag{4·2}$$

because, as in our randomised trial example, the subjects with complete data represent a biased sample. Note that, given our assumption (3·1), (4·2) is implied by the identifiable restriction

$$\text{pr}\,(R_{it} = 1 | R_{i(t-1)} = 1, \bar{W}_{it}) \equiv \text{pr}\,(R_{it} = 1 | R_{i(t-1)} = 1, X_i). \tag{4·3}$$

Now, let

$$U_i(\beta, \alpha) := \bar{\pi}_{iT}^{-1}(\alpha) R_{iT} D_i \varepsilon_i(\beta) \tag{4·4}$$

and let $\tilde{\beta}(\hat{\alpha})$ be the solution to $\sum U_i(\beta, \alpha) = 0$. In the Appendix we prove the following.

PROPOSITION 5. *If* (3·1)–(3·3) *and* (4·1) *hold then Proposition* 4 *remains true with*

$$Q_{it} := q_t(\bar{W}_{it}) := D_i E(\varepsilon_i | \bar{W}_{it}, R_{i(t-1)} = 1)/\bar{\pi}_{it} \tag{4·5}$$

and $U_i(\beta, \alpha)$ *as defined in* (4·4).

5. EFFICIENCY CONSIDERATIONS

For a fixed choice of $d(X_i)$, the asymptotic variance of $\tilde{\beta}(\hat{\alpha}_{\text{op}})$ attains the variance bound $\tau^{-1} \, \text{var}\,(U_i - S_i) \tau^{-1\prime}$ where $\hat{\beta}(\hat{\alpha}_{\text{op}})$ is defined as follows. Given a correctly specified model $\bar{\lambda}_{it}(\alpha)$ satisfying (3·3), let $\bar{\lambda}_{it}(\alpha_{op})$ be the expanded model

$$\operatorname{logit} \bar{\lambda}_{it}(\alpha_{op}) = \operatorname{logit} \bar{\lambda}_{it} + \sigma' q_t(\bar{W}_{it}), \tag{5·1}$$

with $\alpha_{op} := (\alpha', \sigma')'$ and $\alpha_{op,0} := (\alpha_0', \sigma_0')'$. The true value σ_0 of σ in (5·1) equals zero since (3·3) is true. Then $\tilde{\beta}(\hat{\alpha}_{op})$, with $\hat{\alpha}_{op}$ the partial maximum likelihood estimator of α_{op} in model (5·1), attains the bound since $b\, \partial \operatorname{logit} \bar{\lambda}_{it}(\alpha_{op,0})/\partial\alpha_{op}$ equals $\bar{Q}_{it}$ with b the $p \times (v+p)$ matrix $(0, I_{p\times p})$ and $I_{p\times p}$ the $p \times p$ identity matrix. However, $\tilde{\beta}(\hat{\alpha}_{op})$ is not available for data analysis since it depends on the unknown population quantity $q_t(\bar{W}_{it}) := E(\varepsilon_i \mid \bar{W}_{it}, R_{i(t-1)} = 1)/\bar{\pi}_{it}$. Therefore, following Amemiya (1985, p. 186), we shall refer to $\tilde{\beta}(\hat{\alpha}_{op})$ as an infeasible estimator. In § 6, we provide a feasible estimator whose asymptotic variance should nearly attain the bound. The infeasible estimator $\tilde{\beta}(\hat{\alpha}_{op})$ also depends on the choice of $d(X_i)$. In the absence of missing data, the optimal choice for $d(X_i)$ is known to be

$$D_{\text{full},i} := d_{\text{full}}(X_i)' := \partial g(X_i, \beta_0)/\partial\beta \{\operatorname{var}(\varepsilon_i \mid X_i)\}^{-1}$$

(McCullagh, 1983; Chamberlain, 1987). With incomplete data the optimal choice of $d(X_i)$ in the sense of minimising the asymptotic variance of the infeasible estimator $\tilde{\beta}(\hat{\alpha}_{op})$ is

$$D_{op,i} = \{\partial g(X_i, \beta_0/\partial\beta\} \{\operatorname{var}(R_{iT}\bar{\pi}_{iT}^{-1}\varepsilon_i - S_i^* \mid X_i)\}^{-1}, \tag{5·2}$$

where

$$S_i^* = \sum_{t=1}^{T} (R_{it} - \bar{\lambda}_{it} R_{i(t-1)}) \bar{\pi}_{it}^{-1} E(\varepsilon_i \mid \bar{W}_{it}, R_{i(t-1)} = 1).$$

This follows from the following proposition, proved in the Appendix, that states that the asymptotic variance of the infeasible estimator $\tilde{\beta}^*(\hat{\alpha}_{op})$ that uses $D_{op,i}$ in (4·4) and (4·5) attains the semiparametric variance bound.

PROPOSITION 6. *The asymptotic variance of* $n^{\frac{1}{2}}\{\tilde{\beta}^*(\hat{\alpha}_{op}) - \beta_0\}$ *is*

$$\tau_{op}^{-1} := [E\{D_{op,i}\, \partial g(X_i, \beta_0)/\partial\beta'\}]^{-1}.$$

Further τ_{op}^{-1} *is the semiparametric variance bound for regular estimators of* β_0 *in the semiparametric models (a), (b), (c) and (d) defined by the restrictions:*

(*a*): (3·1), (3·2), (4·1);
(*b*): (3·3)–(3·3), (4·1);
(*c*): (3·1), (3·2), (4·1) *and* $\bar{\lambda}_{it}$ *completely known*;
(*d*): (3·1), (3·2), (4·1) *and* (4·3).

Proposition 6 states that, given (3·1) and (3·2), complete or partial knowledge of the probabilities $\bar{\lambda}_{it}$ does not asymptotically provide additional information about β_0. Specifically the bound is the same whether

(*a*) $\bar{\lambda}_{it}$ is completely unknown,
(*b*) $\bar{\lambda}_{it}$ is known up to a vector of unknown parameters,
(*c*) $\bar{\lambda}_{it}$ is completely known, or
(*d*) the data are known to be missing completely at random; i.e. (4·3) and (4·2) are true.

A feasible adaptive estimator whose asymptotic variance may nearly attain the bound is proposed in § 6 below.

When the data are missing completely at random, i.e. (4·3) is also true, the most efficient estimator of β_0 that does not use data on W_{it} is the complete case estimator $\tilde{\beta}_{comp}$ solving $\sum_i R_{iT} D_{\text{full},i}\varepsilon_i(\beta) = 0$, the optimal weighted least squares estimator restricted to subjects with complete data. The infeasible estimator $\tilde{\beta}^*(\hat{\alpha}_{op})$ and its feasible adaptive counterpart given in § 6 both have smaller asymptotic variance than $\hat{\beta}_{\text{comp}}$ because, in contrast to $\hat{\beta}_{\text{comp}}$, they exploit the correlation between $\bar{W}_{iT}$ and Y_i.

If, in addition to (3·1) and (3·2),

$$\text{pr}(R_{it}=1 \mid R_{i(t-1)}=1, \bar{W}_{iT}, Y_i) \equiv \text{pr}(R_{it}=1 \mid R_{i(t-1)}=1, \bar{W}_{it}, Y_i), \tag{5·3}$$

then the data are 'missing at random' in the sense of Rubin (1976). In the proof of Proposition 6 in the Appendix, we show that the semiparametric efficiency bound τ_{opt}^{-1} in models (*a*)–(*d*) of Proposition 6 is unchanged if we impose the additional assumption that the data are 'missing at random' in the sense of Rubin (1976).

When Y_i is Bernoulli, the above results allow us to improve on an estimator recently proposed by Pepe (1992). Since the density of a Bernoulli random variable is completely specified by its mean, (4·1) is a model for the conditional density of Y_i given X_i. In this setting, in her Example 1, Pepe (1992) proposed an estimator $\hat{\beta}_{\text{Pepe}}$ for β_0 in (4·1) based on an 'estimated likelihood function' when Y_i is missing completely at random and $T=2$. Her goal was to extract information about β_0 from the surrogate data $W_i := W_{i1}$ on subjects missing Y_i. However, she noted that $\hat{\beta}_{\text{Pepe}}$ may be less efficient than the complete case estimator $\hat{\beta}_{\text{comp}}$ if the proportion π in the validation sample is small or W_i is not a good predictor of Y_i. In contrast, the asymptotic variance of $\tilde{\beta}^*(\hat{\alpha}_{\text{op}})$ will always be at least as small as the more efficient of $\hat{\beta}_{\text{comp}}$ and $\hat{\beta}_{\text{Pepe}}$, since it equals the semiparametric variance bound.

6. Nearly efficient adaptive estimation

The infeasible estimator $\tilde{\beta}^*(\hat{\alpha}_{\text{op}})$ is not available for data analysis since $q_t(\bar{W}_{it})$ and $d_{\text{op}}(X)$ depend on the unknown probability law generating the data. To obtain a feasible estimator, we require estimates $\hat{q}_t(\bar{W}_{it})$ and $\hat{D}_{\text{op},i}$ of $q_t(\bar{W}_{it})$ and $D_{\text{op},i}$. Given a preliminary inefficient estimator $\tilde{\beta}(\hat{\alpha})$ and estimates $\hat{E}(\varepsilon_i \mid \bar{W}_{it}, R_{i(t-1)}=1)$, we first estimate S_i^* and then obtain $\hat{D}_{\text{op},i}$ from (5·2) and $\hat{q}_t(\bar{W}_{it})$ from (4·5) with $\hat{D}_{\text{op},i}$ substituted for D_i.

To estimate $E(\varepsilon_i \mid \bar{W}_{it}, R_{i(t-1)}=1)$, we cannot simply regress the residuals $\varepsilon_i\{\tilde{\beta}(\hat{\alpha})\}$ on functions of $\bar{W}_{it}$ among subjects observed at the end of follow-up since the missingness mechanism (3·1) does not imply that

$$E(\varepsilon_i \mid \bar{W}_{it}, R_{i(t-1)}=1) \equiv E(\varepsilon_i \mid \bar{W}_{it}, R_{iT}=1). \tag{6·1}$$

However, it can be verified that the left-hand side of (6·1) is equal to

$$E\{\bar{\pi}_{iT}^{-1}\bar{\pi}_{i(t-1)}\varepsilon_i(\beta_0) \mid R_{iT}=1, \bar{W}_{it}\}\ \text{pr}(R_{iT}=1 \mid R_{i(t-1)}=1, \bar{W}_{it}).$$

Thus, consider flexible regression models

$$E\{\bar{\pi}_{iT}^{-1}\bar{\pi}_{i(t-1)}\varepsilon_i(\beta_0) \mid R_{iT}=1, \bar{W}_{it}\} \equiv m_t(\delta_t, \bar{W}_{it}), \tag{6·2}$$

$$\text{pr}(R_{iT}=1 \mid R_{i(t-1)}=1, \bar{W}_{it}) \equiv l_t(\chi_t, \bar{W}_{it}), \tag{6·3}$$

depending on finite dimensional parameters δ_t and χ_t, with $m_t(.,.)$ and $l_t(.,.)$ regression functions chosen by the investigator. Often $l_t(.,.)$ will be the logistic function. Given preliminary estimates $\tilde{\beta}(\hat{\alpha})$ of β_0 and $\hat{\alpha}$ of α_0, let $\hat{\delta}_t$ be the maximum likelihood estimator of δ_t and let $\hat{\chi}_t$ be the least squares estimator of χ_t in the regression of $\bar{\pi}_{iT}^{-1}(\hat{\alpha})\bar{\pi}_{i(t-1)}(\hat{\alpha})\varepsilon_i\{\tilde{\beta}(\hat{\alpha})\}$ on $\bar{W}_{it}$ among subjects observed at the end of follow-up. The estimate of S_i^* is given by

$$\hat{S}_i^* = \sum \{R_{it} - \bar{\lambda}_{it}(\hat{\alpha})R_{i(t-1)}\} m_t(\hat{\delta}_t, \bar{W}_{it}) l_t(\hat{\chi}_t, \bar{W}_{it}).$$

Next, we estimate var $(\bar{\pi}_{iT}^{-1}\varepsilon_i - S_i^* \mid X_i)$. Consider the possibly nonlinear regression model

$$E\{(\bar{\pi}_{iT}^{-1}R_{iT}\varepsilon_i - S_i^*)^2 \mid X_i\} \equiv h(\psi, X_i) \tag{6·4}$$

depending on a parameter ψ and a regression function $h(.,.)$ chosen by the investigator. We estimate ψ by $\hat{\psi}$, the possibly nonlinear least squares regression of $[\bar{\pi}_{iT}^{-1}R_{iT}\varepsilon_i\{\tilde{\beta}(\hat{\alpha})\} - \hat{S}_i^*]^2$ on X_i $(i = 1, \ldots, n)$. The estimate $\hat{D}_{\text{op},i}$ is given by

$$\hat{D}_{\text{op}}(X_i) = [\partial g\{X_i, \tilde{\beta}(\hat{\alpha})\}/\partial\beta']h(\hat{\psi}, X_i).$$

The estimate $\hat{q}_t(\bar{W}_{it})$ is given by

$$\hat{q}_t(\bar{W}_{it}) = \hat{D}_{\text{op}}(X_i)\bar{\pi}_{it}^{-1}(\hat{\alpha})m_t(\hat{\delta}_t, \bar{W}_{it})l_t(\hat{\chi}_t, \bar{W}_{it}).$$

Let $\tilde{\alpha}_{\text{op}}$ denote the partial maximum likelihood estimator of α_{op} in model (5·1) with $\hat{q}_t(\bar{W}_{it})$ replacing $q_t(\bar{W}_{it})$. Finally $\hat{\beta}^*(\tilde{\alpha}_{\text{op}})$ is the estimate that uses $\hat{q}_t(\bar{W}_{it})$ and $\hat{D}_{\text{op}}(X_i)$.

It is standard to show that, when (6·2)–(6·4) are correctly specified, $\hat{\beta}^*(\tilde{\alpha}_{\text{op}})$ has the same limiting distribution as $\tilde{\beta}^*(\hat{\alpha}_{\text{op}})$ and thus will attain the semiparametric variance bound (Robins et al., 1992). Furthermore, $\hat{\beta}^*(\tilde{\alpha}_{\text{op}})$ will be consistent for estimating β_0 even when, as will usually be the case, (6·2), (6·3) or (6·4) is false. Nevertheless, provided that models (6·2)–(6·4) are flexible, i.e. highly parametrised, the asymptotic variance of $\hat{\beta}^*(\tilde{\alpha}_{\text{op}})$ should nearly attain the semiparametric variance bound. A consistent estimate of its variance that is robust to misspecification of either (6·2), (6·3) or (6·4) is as given in Proposition 5 with $\hat{D}_{\text{op},i}$ and $\tilde{\alpha}_{\text{op}}$ substituted for $D_{\text{op},i}$ and $\hat{\alpha}$.

Robins, Rotnitzky & Zhao (1995) have constructed analogous adaptive estimators in a multivariate generalisation of our model. For completeness we have included their construction.

7. Further considerations

The methods presented in this paper can be extended to a variety of missing data problems. Robins et al. (1995) and Robins & Rotnitzky (1995) extend our results to the analysis of repeated outcomes when missingness is independent of unobserved outcomes given past observations. Robins et al. (1992) and Robins et al. (1994) show how to construct locally efficient inverse probability weighted estimators in an arbitrary semiparametric model in which the data are missing at random and the probability of observing complete data is bounded away from zero.

Acknowledgement

Support for this research was provided in part by grants from the National Institutes of Health. Andrea Rotnitzky was additionally supported in part by a Mellon Foundation Faculty Development Award.

Appendix

Proofs

We sketch a proof of Propositions 5 and 6 only, since Propositions 1, 2, 3 and 4 are special cases. Throughout, the parts of Proposition 5 corresponding to Propositions 1, 2 and 3 will be referred to as the analogues of these propositions. Under the following regularity conditions, a rigorous proof of Proposition 5 can be constructed in a fashion entirely analogous to the proof of Proposition (6.1) of Robins et al. (1994). A sketch of this proof is given below.

Regularity conditions. Let $H_i(\gamma)=\{D_i\varepsilon_i(\beta)', S_{\alpha,i}(\alpha)'\}'$ with $\gamma':=(\beta',\alpha')$. We assume:

(i) β and α lie in the interior of compact sets β^* and α^*;

(ii) $(Y_i, \bar{W}_{iT})$ $(i=1,\ldots,n)$ are independent and identically distributed;

(iii) $\lambda_{it}(\alpha)>c>0$ for all $\alpha\in\alpha^*$ $(t=1,\ldots,T)$ for some c;

(iv) $E\{H_i(\gamma)\}=0 \Leftrightarrow \gamma=\gamma_0$;

(v) var $\{H_i(\gamma_0)\}$ is finite and positive definite;

(vi) $E\{\partial H_i(\gamma_0)/\partial\gamma'\}$ exists and is invertible;

(vii)

$$E\left\{\sup_{\gamma\in\gamma}\|H_i(\gamma)\|\right\},\quad E\left\{\sup_{\gamma\in\gamma}\|\partial H_i(\gamma)/\partial\gamma'\|\right\},\quad E\left\{\sup_{\gamma\in\gamma}\|H_i(\gamma)H_i(\gamma)'\|\right\}$$

are all finite where $\|A\|:=\{\sum_{ij}A_{ij}^2\}^{1/2}$ for any matrix A with elements A_{ij}, and γ^* is the Cartesian product of α^* and β^*.

Sketch of proof of Proposition 5. We first show that $E\{U_i(\beta_0,\alpha_0)\}=0$. Consider the identity

$$R_{iT}\bar{\pi}_{iT}^{-1}=1+\sum_{t=1}^{T}(R_{it}-\bar{\lambda}_{it}R_{i(t-1)})\bar{\pi}_{it}^{-1}.$$

Multiplying both sides by $D_i\varepsilon_i(\beta)$ we get

$$U_i(\beta,\alpha_0)=D_i\varepsilon_i(\beta)+\sum(R_{it}-\bar{\lambda}_{it}R_{i(t-1)})\bar{\pi}_{it}^{-1}D_i\varepsilon_i(\beta).$$

But

$$E\{D_i\varepsilon_i(\beta_0)\}=E[D_iE\{\varepsilon_i(\beta_0)|X_i\}]=0,$$

$$E\{(R_{it}-\bar{\lambda}_{it}R_{i(t-1)})\bar{\pi}_{it}^{-1}D_i\varepsilon_i(\beta_0)\}=E\{E(R_{it}-\bar{\lambda}_{it}R_{i(t-1)}|R_{i(t-1)},\bar{W}_{it},Y_i)\bar{\pi}_{it}^{-1}D_i\varepsilon_i(\beta_0)\}=0,$$

because, by (3·1) and the monotonicity assumption, $\bar{\lambda}_{it}R_{i(t-1)}=E(R_{it}|R_{i(t-1)},\bar{W}_{it},Y_i)$.

We next prove the analogues of parts (i) and (ii) of Proposition 1. Robins et al. (1994, proof of Proposition 6.1) show that the regularity conditions guarantee the validity of the standard Taylor expansions

$$n^{\frac{1}{2}}(\hat{\alpha}-\alpha_0)=-E\left(\frac{\partial S_{\alpha i}}{\partial\alpha'}\right)^{-1}n^{-\frac{1}{2}}S_\alpha(\alpha_0)+o_p(1),\tag{A·1}$$

$$n^{\frac{1}{2}}\{\tilde{\beta}(\hat{\alpha})-\beta_0\}=-E\left(\frac{\partial U_i}{\partial\beta'}\right)^{-1}\left\{n^{-\frac{1}{2}}U(\beta_0,\alpha_0)-E\left(\frac{\partial U_i}{\partial\alpha'}\right)E\left(\frac{\partial S_{\alpha i}}{\partial\alpha'}\right)^{-1}n^{-\frac{1}{2}}S_\alpha(\alpha_0)\right\}+o_p(1),\tag{A·2}$$

$$n^{\frac{1}{2}}\{\tilde{\beta}(\alpha_0)-\beta_0\}=-E\left(\frac{\partial U_i}{\partial\beta'}\right)^{-1}n^{-\frac{1}{2}}U(\beta_0,\alpha_0)+o_p(1),\tag{A·3}$$

where

$$S_\alpha(\alpha):=\sum S_{\alpha i}(\alpha),\quad S_{\alpha i}:=S_{\alpha i}(\alpha_0),\quad U(\beta,\alpha):=\sum U_i(\beta,\alpha),\quad U_i:=U_i(\beta_0,\alpha_0),$$
$$\partial S_{\alpha i}/\partial\alpha':=\partial S_{\alpha i}(\alpha_0)/\partial\alpha',\quad \partial U_i/\partial\beta':=\partial U_i(\beta_0)/\partial\beta'.$$

The asymptotic distribution of $n^{\frac{1}{2}}(\tilde{\beta}(\alpha_0)-\beta_0\}$ now follows from (A·3) using the Central Limit Theorem. To obtain the limiting distribution of $n^{\frac{1}{2}}\{\tilde{\beta}(\hat{\alpha})-\beta_0\}$, note that, by differentiating $E_{\beta_0,\alpha}\{U_i(\beta_0,\alpha)\}=0$ with respect to α and evaluating at α_0, it follows that $E(\partial U_i/\partial\alpha')=-E(U_iS_{\alpha i}')$ (Pierce, 1982). Similarly, $-E(\partial S_{\alpha i}/\partial\alpha)=\text{var}\,(S_{\alpha i})$. Substituting in (A·1) and (A·2) we obtain,

$$n^{\frac{1}{2}}\{\tilde{\beta}(\hat{\alpha})-\beta_0\}=-\tau^{-1}n^{-\frac{1}{2}}\sum\text{resid}\,(U_i,S_{\alpha i})+o_p(1)\tag{A·4}$$

and the asymptotic distribution of $n^{\frac{1}{2}}\{\tilde{\beta}(\hat{\alpha})-\beta_0\}$ follows from (A·4) and the Central Limit Theorem. The consistency of $\tilde{\tau}$, $\tilde{V}$ and $\tilde{C}$ is a consequence of the Law of Large Numbers.

The analogue of parts (i) and (ii) of Proposition 2 is proved as in § 2·2 except that

$$G = \left\{ \sum_{t=1}^{T} (R_{it} - \bar{\lambda}_{it} R_{i(t-1)}) g_t(\bar{W}_{it}) \right\}$$

and $\operatorname{cov}\{(U_i - S_i), (S_i - G_i)\}$ is replaced by

$$\tfrac{1}{2}[E\{(U_i - S_i)(S_i - G_i)'\} + E\{(S_i - G_i)(U_i - S_i)'\}].$$

Part (iii) is a corollary of Proposition 6.

To prove the analogue of Proposition 3 we note that $S_{\alpha^{(j)},i}$ is the first $v^{(j)}$ components of the $v^{(j+1)}$ vector $S_{\alpha^{(j+1)},i}$. By standard least squares theory, the variance of a residual from a population regression does not increase as the number of regressors increases. Hence if $C^{(j)}$ denotes the variance of the residual from a population regression of U_i on $S_{\alpha^{(j)},i}$ for $\alpha^{(j)}$ in the jth model, we have $C^{(j)} \geqslant C^{(j+1)}$. Thus,

$$\operatorname{var}^A [n^{\frac{1}{2}}\{\tilde{\beta}(\hat{\alpha}^{(j)}) - \beta_0\}] = \tau^{-1} C^{(j)} \tau^{-1'} \geqslant \tau^{-1} C^{(j+1)} \tau^{-1'} = \operatorname{var}^A [n^{\frac{1}{2}}\{\tilde{\beta}(\hat{\alpha}^{(j+1)}) - \beta_0\}],$$

since $\tau = E(\partial U_i/\partial \beta')$ does not depend on model (3·3). □

Proof of Proposition 6. Our proof uses the following.

Step 1. We first show that the semiparametric variance bound in models (a)–(d) are equal and also equal the bound in models (a^*)–(d^*) corresponding to models (a)–(d) but with (5·3) also imposed so that the data are missing at random.

Step 2. We show that the asymptotic variance of $n^{\frac{1}{2}}\{\tilde{\beta}^*(\hat{\alpha}_{\text{op}}) - \beta_0\}$ is the semiparametric variance bound in model c^* and hence, in view of Step 1, it is the asymptotic variance bound in models (a)–(d) and (a^*)–(d^*).

Proof. Robins et al. (1994, Proposition 8.1e1) state that the semiparametric variance bound does not depend on knowledge of the missingness process when the data are missing at random and thus implies that models (a^*)–(d^*) have identical bounds: see Remark 1 following Proposition 8.1 of Robins et al. (1994).

We next show that the bounds in models (a) and (a^*) are identical. Let Z_i be the observed data vector for the ith subject, that is, $Z_i = (\bar{W}'_{iT}, Y_i, R'_i)'$ if $(\bar{W}'_{iT}, Y_i, R'_i)'$ is fully observed, and $Z_i = (\bar{W}'_{it}, R'_i)'$ if subject i is lost to follow-up at time t. Since (5·3) is nonidentifiable in the sense that it may be true whatever be the distribution of the observables Z_i, the allowable distributions for Z_i are the same under models (a) and (a^*). Hence by definition, models (a) and (a^*) are the same semiparametric model for the observables and thus have the same efficient score and variance bound. The same argument shows that the bounds are identical in models (b) and (b^*), (c) and (c^*), and (d) and (d^*).

We next turn to Step 2. Robins et al. (1994, § 7.1) show that their Propositions (3·2) and (4·2) imply that, in their notation, $[\operatorname{var}\{D(\alpha_0, h_{\text{eff}}, \phi^{h_{\text{eff}}})\}]^{-1}$ attains the semiparametric variance bound in the model (c^*) that assumes the data are missing at random and the missingness probabilities are known. Here, $h_{\text{eff}}(X^*)$ is the unique solution to their equation (4.4). By rewriting this result in the notation of this paper, we will show that Step 2 follows as a corollary. Let $A \to B$ mean that the symbol A of Robins et al. (1994, § 7.1) is identified with the symbol B in the current paper. Then we have $\alpha \to \beta$, $K \to T$, $h(X_i^*) \to d(X_i)$, $I(R=1) \to R_{iT}$, $\bar{\pi}_{ik} \to \bar{\pi}_{ik}$, $\pi_{it} \to \lambda_{it}$, $\varepsilon \to \varepsilon$, $\bar{L}_{ik} \to \bar{W}_{ik}$, $R_{ik} \to R_{ik}$. With these identifications, we find that $A_i(\phi) \to S_i(d)$, where $A(\phi)$ and ϕ^h are defined by their equations (7·5) and (7·6) and $S_i(d) := S_i$ with S_i defined in our Proposition 4 using the $\bar{Q}_{it}$ in equation (4·5). Thus

$$D(\alpha_0, h, \phi^h) \to \bar{\pi}_{iT}^{-1} R_{iT} d(X_i)\varepsilon_i - S_i(d) =: \Omega(d)$$

and equation (4·4) $\to$

$$d(X_i) = \left(\frac{\partial g(X_i; \beta_0)}{\partial \beta} + E[r^*\{d(X_i)\varepsilon_i\}\varepsilon_i | X_i] \right) \left\{ \operatorname{var}\left(\frac{\varepsilon_i}{\pi_{iT}} \right) \right\}^{-1}, \tag{A·5}$$

where

$$r^*(B_i) := \sum_{t=1}^{T} (1 - \lambda_{it}) \bar{\pi}_{it}^{-1} E(B_i \mid \bar{W}_{ik}, R_{i(k-1)} = 1),$$

$r(B_i) \rightarrow r^*(B_i)$ and $r(B_i)$ is given by their equation (7.9). It is straightforward to show by direct substitution into (A·5) that $d_{\rm op}(X_i)$ solves (A·5) and that the asymptotic variance of $\tilde{\beta}^*(\hat{\alpha}_{\rm op})$ is $[\mathrm{var}\{\Omega(d_{\rm op})\}]^{-1}$, proving Proposition 6. □

REFERENCES

AMEMIYA, T. (1985). *Advanced Econometrics.* Cambridge, MA: Harvard University Press.

BEGUN, J. M., HALL, W. J., HUANG, W. M. & WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11**, 432–52.

CHAMBERLAIN, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *J. Economet.* **34**, 305–24.

FLANDERS, W. D. & GREENLAND, S. (1991). Analytic methods for two stage case-control studies and other stratified designs. *Statist. Med.* **10**, 739–47.

GILL, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part 1). *Scand. J. Statist.* **16**, 97–128.

HEYTING, A., TOLBOOM, J. T. B. M. & ESSERS, J. G. A. (1992). Statistical handling of drop-outs in longitudinal clinical trials. *Statist. Med.* **11**, 2043–62.

HORVITZ, D. G. & THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.* **47**, 663–85.

HUBER, P. (1981). *Robust Statistics.* New York: John Wiley.

HUBER, P. (1985). Projection pursuit. *Ann. Statist.* **13**, 435–74.

KALBFLEISCH, J. D. & LAWLESS, J. F. (1988). Likelihood analysis of multi-stage models for disease incidence and mortality. *Statist. Med.* **7**, 149–60.

KIEFER, J. & WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Ann. Math. Statist.* **27**, 887–906.

MANSKI, C. F. & LERMAN, S. (1977). The estimation of choice probabilities from choice-based samples. *Econometrica* **45**, 1977–88.

MCCULLAGH, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11**, 59–67.

NEWEY, W. K. (1995). The asymptotic variance of semiparametric estimators. *Econometrica.* To appear.

PEPE, M. S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika* **79**, 355–66.

PIERCE, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *Ann. Statist.* **10**, 475–8.

PRENTICE, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statist. Med.* **8**, 431–40.

ROBINS, J. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* **79**, 321–34.

ROBINS, J. M., MARK, S. D. & NEWEY, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrika* **48**, 479–95.

ROBINS, J. M. & ROTNITZKY, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology—Methodological Issues*, Ed. N. Jewell, K. Dietz and V. Farewell, pp. 297–331. Boston: Birkhäuser.

ROBINS, J. M. & ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Statist. Assoc.* **90**, 122–9

ROBINS, J. M., ROTNITZKY, A. & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–86.

ROBINS, J. M., ROTNITZKY, A. & ZHAO, L.-P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Statist. Assoc.* **90**, 106–21.

ROSENBAUM, P. R. (1987). Model-based direct adjustment. *J. Am. Statist. Assoc.* **82**, 387–94.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–92.

ZHAO, L. P. & LIPSITZ, S. (1992). Designs and analysis of two-stage studies. *Statist. Med.* **11**, 769–82.

[*Received June* 1991. *Revised January* 1995]