Biometrika (2012), **99**, 2, *pp.* 439–456
doi: 10.1093/biomet/ass013
Advance Access publication 29 April 2012

Improved double-robust estimation in missing data and causal inference models

By ANDREA ROTNITZKY

Di Tella University, Saenz Valiente 1010, Buenos Aires 14281, Argentina
arotnitzky@utdt.edu

QUANHONG LEI

Adheris, Inc., One Van de Graaff Drive, Burlington, Massachusetts 01803, U.S.A.
leiqlei@gmail.com

MARIELA SUED

Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Guiraldes 2160, Buenos Aires 1428, Argentina
marielasued@gmail.com

and JAMES M. ROBINS

Harvard School of Public Health, 655 Huntington Ave., Boston, Massachusetts 02115, U.S.A.
robins@hsph.harvad.edu

Summary

Recently proposed double-robust estimators for a population mean from incomplete data and for a finite number of counterfactual means can have much higher efficiency than the usual double-robust estimators under misspecification of the outcome model. In this paper, we derive a new class of double-robust estimators for the parameters of regression models with incomplete cross-sectional or longitudinal data, and of marginal structural mean models for cross-sectional data with similar efficiency properties. Unlike the recent proposals, our estimators solve outcome regression estimating equations. In a simulation study, the new estimator shows improvements in variance relative to the standard double-robust estimator that are in agreement with those suggested by asymptotic theory.

Some key words: Drop-out; Marginal structural model; Missing at random.

1. Introduction

In a missing data model, an estimator is double-robust if it is consistent when either a model for the missingness mechanism or a model for full-data law is correctly specified. In a causal inference model, an estimator is double-robust if it is consistent when either a model for the treatment assignment mechanism or a model for the counterfactual data distribution is correctly specified.

Scharfstein et al. (1999) noted that an estimator originally developed and identified as the locally efficient estimator in the class of augmented inverse probability weighted estimators in

missing data models in Robins et al. (1994), was double-robust. Since then, many estimators with the double-robust property have been proposed, several of which were recently reviewed by Kang & Schafer (2007) and the discussants of that paper, for the special case of estimating a population mean and the causal effect of a binary treatment.

Rubin & van der Laan (2008) noted that the locally efficient estimator of Robins et al. (1994) can be quite inefficient if the model for the full-data distribution is incorrectly specified. To remedy this, they described a new general approach yielding locally efficient estimators with desirable efficiency properties when the full-data model is incorrectly specified. Tan (2008) and Cao et al. (2009) demonstrated that for estimating a population mean with missing data and unknown missingness probabilities, a particular form of the Rubin and van der Laan procedure yields double-robust estimators. In a recent paper, Tan (2010a) combines that procedure with restricted empirical maximum likelihood estimation to derive new double-robust estimators of a population mean with missing data and of population average treatment effects that have the efficiency properties of the Rubin and van der Laan estimator. A property of the procedure in Tan (2010a) not satisfied by the proposals of Tan (2008) and Cao et al. (2009) is that means are estimated as weighted averages with positive weights. Thus, estimated means always fall in the parameter space and in the range of observed outcomes.

In this paper, we describe a new general approach to constructing locally efficient double robust estimators for the parameters of regression models with outcomes missing at random and for parameters of marginal structural mean models for point exposure studies with continuous or discrete exposures that have the advantageous efficiency properties of the Rubin and van der Laan procedure. For the special case of a population mean, our estimators do not reduce to any of the earlier estimators of Tan (2008, 2010a) and Cao et al. (2009). In fact, unlike these other proposals, our estimators solve outcome regression estimating equations. These are equations identical to the ordinary weighted least squares estimating equations but with the missing outcome or counterfactual response replaced by an estimate of its conditional expectation given baseline covariates. As such, unlike augmented inverse probability weighted estimating equations, our equations always have a solution, as long as their full-data analogues also have a solution and the estimated conditional expectation falls in the sample space of the outcome. In particular, like Tan (2010a), our estimators of a population mean always fall in the parameter range.

Several proposals exist for constructing locally efficient double-robust estimators that solve outcome regression estimating equations. For regression models with missing data and for marginal structural models, these include the procedures in Bang & Robins (2005) and the targeted maximum likelihood methodology (van der Laan & Rubin, 2006; van der Laan, 2010). For a population mean, Kang & Schafer (2007, Equation (10)) described a so-called double-robust weighted least squares outcome regression estimator. These authors reported a simulation study in which this estimator performed better than the Bang and Robins estimator. These approaches have not yet been adapted to provide estimators with the improved efficiency properties of Rubin & van der Laan (2008).

The procedure described here is essentially a generalization to the regression setting of Kang and Shafer's weighted least squares outcome regression estimator of a population mean. The key innovation is that the weights depend on an augmented model for the missingness or treatment probability which incorporates covariates constructed so as to ensure the advantageous efficiency properties of the Rubin & van der Laan (2008) procedure. Our approach exploits the counterintuitive fact that the efficiency of augmented inverse probability weighted estimators improves as the dimension of the model under which the missingness or treatment probabilities are estimated increases.

2. Cross-sectional studies with missing data

2·1. *Existing double-robust estimators*

Consider a study in which the intended data, (Y_i, L_i) $(i = 1, \ldots, n)$, are independent and identically distributed across i, where Y_i is a scalar outcome and L_i is a vector of additional variables. Assume L_i is observed but Y_i is missing in a subsample. Let $A_i = 1$ if Y_i is observed and $A_i = 0$ otherwise. In this section, we consider estimation of the unknown $p \times 1$ parameter vector β_0 of the regression model

$$E(Y \mid Z) = h(Z; \beta_0), \tag{1}$$

where $h(\cdot\,; \cdot)$ is a known smooth function and Z is a subset of the components of L, possibly empty, from independent and identically distributed copies $(A_i, A_i Y_i, L_i^{\mathrm{T}})$ $(i = 1, \ldots, n)$, of (A, AY, L^{T}). We make the missing at random assumption (Rubin, 1976) that $f(A \mid Y, L) = f(A \mid L)$ and the positivity assumption that $\mathrm{pr}(A = 1 \mid Y, L) > 0$. Under these assumptions, β_0 solves for any $p \times 1$ function $b(\cdot)$, the population moment equation

$$E[b(Z)\{E(Y \mid L, A = 1) - h(Z; \beta)\}] = 0. \tag{2}$$

Equation (2) has motivated the so-called outcome regression estimator $\hat{\beta}_{\mathrm{reg}}(\hat{\eta})$. To compute it, one first estimates the unknown parameter η_0 of a working outcome regression model for the respondents

$$E(Y \mid L, A = 1) = m(L; \eta_0), \tag{3}$$

where $m(\cdot\,; \cdot)$ is a known smooth function, by some $\hat{\eta}$ using the units with observed Y. Next, one computes $\hat{\beta}_{\mathrm{reg}}(\hat{\eta})$ where, by definition, for any η, $\hat{\beta}_{\mathrm{reg}}(\eta)$ solves

$$E_n[b(Z)\{m(L; \eta) - h(Z; \beta)\}] = 0, \tag{4}$$

with $b(\cdot)$ a $p \times 1$ user-specified vector function. In what follows, $E_n(\cdot)$ and $E_n(\cdot \mid \cdot)$ stand for the empirical mean and conditional mean operator, i.e., $E_n(U) \equiv n^{-1} \sum_{i=1}^n U_i$, and $E_n(U \mid A = 1, V) \equiv \{\sum_{i=1}^n I(A_i = 1, V_i = V) U_i\} / \{\sum_{i=1}^n I(A_i = 1, V_i = V)\}$. Under regularity conditions which include the requirement that (2) has a unique solution, the estimator $\hat{\beta}_{\mathrm{reg}}(\hat{\eta})$ is consistent for β_0 and asymptotically normal if model (3) is correctly specified provided $\hat{\eta}$ is consistent for η_0 and asymptotically normal. Note that (4) is the same as the weighted least squares estimating equations $E_n[b(Z)\{Y - h(Z; \beta)\}] = 0$ that one would ordinarily use in the absence of missing data, for example, with $b(Z) = (1, Z^{\mathrm{T}})^{\mathrm{T}}$ if $h(Z; \beta) = Z^{\mathrm{T}}\beta$ or $h(Z; \beta) = \mathrm{expit}(Z^{\mathrm{T}}\beta)$ where $\mathrm{expit}(u) = \{1 + \exp(-u)\}^{-1}$, but with the outcome Y replaced by $m(L; \eta)$. Then, if the range of $m(\cdot\,; \cdot)$ falls in the sample space of Y, equation (4) will ordinarily have a solution.

The moment equation (2) can be re-written as

$$E[b(Z)\omega\{Y - h(Z; \beta)\}] = 0,$$

where $\omega = A f(A \mid L)^{-1}$. This formulation has motivated the so-called augmented inverse probability weighted estimators $\hat{\beta}_{g,\mathrm{aipw}}$. To compute them, one first posits a parametric model for the missingness probability

$$f(A \mid L) = f(A \mid L; \alpha_0) \tag{5}$$

for instance, $f(A \mid L; \alpha) = \mathrm{expit}(\alpha^{\mathrm{T}}\tilde{L})^A \{1 - \mathrm{expit}(\alpha^{\mathrm{T}}\tilde{L})\}^{1-A}$ where $\tilde{L}^{\mathrm{T}} = (1, L^{\mathrm{T}})$, and estimates α_0 by its maximum likelihood estimator $\hat{\alpha}$. Next, one computes $\hat{\beta}_{g,\mathrm{aipw}}$ solving, with α evaluated

at $\hat{\alpha}$, the equations

$$E_n[\omega(\alpha)b(Z)\{Y - h(Z;\beta)\}] - E_n[g(A,L) - E_\alpha\{g(A,L) \mid L\}] = 0, \tag{6}$$

where $g(\cdot)$ and $b(\cdot)$ are user-specified $p \times 1$ vector functions, $\omega(\alpha) = Af(A \mid L;\alpha)^{-1}$ and $E_\alpha\{g(A,L) \mid L\}$ is the conditional expectation of $g(A,L)$ given L under $f(A \mid L;\alpha)$. When $g = 0$, $\hat{\beta}_{g,\text{aipw}}$, throughout denoted by $\hat{\beta}_{\text{ipw}}$, is called an inverse probability weighted estimator. If model (5) is correctly specified, then under regularity conditions, $\hat{\beta}_{g,\text{aipw}}$ is consistent and asymptotically normal with asymptotic variance no smaller than that of $\hat{\beta}_{g_{\text{opt}},\text{aipw}}$ where $g^b_{\text{opt}}(A,L) = \omega b(Z)\{E(Y \mid A,L) - h(Z;\beta_0)\}$. This motivates estimating β_0 by $\hat{\beta}(\hat{\eta},\hat{\alpha})$ where, for a given $b(\cdot)$ and each (η,α), $\hat{\beta}(\eta,\alpha)$ solves

$$E_n[\omega(\alpha)b(Z)\{Y - h(Z;\beta)\}] - E_n[g^b_{\eta,\alpha,\beta}(A,L) - E_\alpha\{g^b_{\eta,\alpha,\beta}(A,L) \mid L\}] = 0, \tag{7}$$

with $g^b_{\eta,\alpha,\beta}(A,L) = \omega(\alpha)b(Z)\{m(L;\eta) - h(Z;\beta)\}$ (Robins & Rotnitzky, 1995).

Regardless of the validity of the outcome regression model (3), if the missingness model (5) is correct no variance correction due to estimation of η is needed, i.e.,

$$\sqrt{n}\{\hat{\beta}(\hat{\eta},\hat{\alpha}) - \hat{\beta}(\eta^*,\hat{\alpha})\} = o_p(1), \tag{8}$$

where η^* is the probability limit of $\hat{\eta}$. When both the outcome regression and the missingness models are correct, no adjustment due to estimation of α on the asymptotic variance of $\hat{\beta}(\eta_0,\hat{\alpha})$ is needed, i.e., $\sqrt{n}\{\hat{\beta}(\eta_0,\hat{\alpha}) - \hat{\beta}(\eta_0,\alpha_0)\} = o_p(1)$.

Robins & Rotnitzky (1995) noted that the estimator $\hat{\beta}(\hat{\eta},\hat{\alpha})$ satisfies:

Property 1. if the missingness model is correct, it is consistent for β_0 and asymptotically normal;

Property 2. if both the outcome regression and missingness models are correct, it has asymptotic variance equal to the smallest asymptotic variance of all estimators $\hat{\beta}_{g,\text{aipw}}$ that use the same b.

Unlike the outcome regression estimator $\hat{\beta}_{\text{reg}}(\eta)$, the estimator $\hat{\beta}(\hat{\eta},\hat{\alpha})$ may fall outside the range of plausible values for β. For example, in the absence of Z and when $h(Z;\beta) = \beta$, (7) is equivalent to

$$\beta = E_n[\omega(\alpha)\{Y - m(L;\eta)\}] + E_n\{m(L;\eta)\}. \tag{9}$$

If Y is binary and $m(L;\eta) = \text{expit}(\eta^{\mathrm{T}}\tilde{L})$, then $0 < E_n\{m(L;\hat{\eta})\} < 1$. However, $|E_n[\omega(\hat{\alpha})\{Y - m(L;\hat{\eta})\}]|$ may be very large if few weights $\omega_i(\hat{\alpha})$ are exceedingly large, as is the case when few units with $A_i = 1$ have relatively very small estimated values $f(A_i \mid L_i;\hat{\alpha})$. Another manifestation of this problem is in estimating equations with no solution. For instance, if we parameterize $E(Y) = \text{expit}(\beta_0)$, (9) with β replaced by $\text{expit}(\beta)$ has no solution if the right-hand side of (9) is outside the interval $(0, 1)$.

Scharfstein et al. (1999) noted that $\hat{\beta}(\hat{\eta},\hat{\alpha})$ is double-robust: it is consistent for β_0 and asymptotically normal provided either model (5) or model (3) is correct. The estimator $\hat{\beta}(\hat{\eta},\hat{\alpha})$ with $\hat{\eta}$ the ordinary or iteratively reweighted least squares estimator based on units with observed Y is known as the standard double-robust estimator. Several authors (Robins & Wang, 2000; Kang & Schafer, 2007; Rubin & van der Laan, 2008) have noted that standard double-robust estimators may have substantial bias and large variance, even under correct specification of the

missingness model, if the estimated missingness probabilities are highly variable and/or the outcome regression model is misspecified. Alternative double-robust estimators have been recently developed to address these problems (van der Laan & Rubin, 2006; van der Laan, 2010; Tan, 2006, 2007, 2008, 2010a, 2010b; Robins et al., 2007; Cao et al., 2009). In particular, Tan (2008, 2010a) and Cao et al. (2009) derived double-robust estimators of $E(Y) = \beta_0$, i.e., in the special case in which Z is absent and $h(Z; \beta) = \beta$, which satisfy Properties 1 and 2 and have the enhanced efficiency benefit that:

Property 3. if the missingness model is correct, they have asymptotic variance smaller than or equal to the asymptotic variance of any estimator $\hat{\beta}(\eta, \hat{\alpha})$ with η fixed but arbitrary even if the outcome regression model is incorrect.

The estimator of Tan (2008) agrees with that of Cao et al. (2009) when $m(L; \eta)$ is linear in η but otherwise it has a subtle difference that ensures that, when the missingness model is correct, it also has asymptotic variance no larger than that of any estimator computed like $\hat{\beta}(\eta, \hat{\alpha})$ but with $m(L; \eta)$ replaced by $c_1 + c_2 m(L; \eta)$ for any c_1 and c_2. Both proposals adapt a particular version of an estimator proposed by Rubin & van der Laan (2008) to the setting in which the missingness probabilities are unknown and estimated under a parametric model. Under both proposals, the estimators are of the form $\hat{\beta}(\tilde{\eta}, \hat{\alpha})$, where $\tilde{\eta}$ minimizes $\hat{\sigma}^2(\eta)$, an adequately chosen consistent estimator of the asymptotic variance $\sigma^2(\eta)$ of $\hat{\beta}(\eta, \hat{\alpha})$. A drawback of these approaches is that they are based on solving equations of the form (7), which may not have a solution or may yield estimators that fall outside the parameter space. To remedy this Tan (2010a) derived an estimator that maximizes a constrained empirical likelihood. A clever choice of constraints combined with a calibration of a linear model for the missingness probabilities ensures that the estimator is double-robust, is efficient in the aforementioned larger class of estimators considered by Tan (2008), satisfies Properties 1–3 and is a weighted average of the observed outcomes.

In the next section, we introduce a new class of estimators $\hat{\beta}_{\rm dr}$ of parameters of regression models with the following properties:

Property 4. they are double-robust, i.e., consistent for β_0 and asymptotically normal if either model (3) or (5) is correct;

Property 5. they satisfy Properties 1 and 2;

Property 6. they solve an outcome regression estimating equation;

Property 7. given user-specified real-valued functions $\phi_1(\cdot), \ldots, \phi_K(\cdot)$, each $\phi_k(\hat{\beta}_{\rm dr})$ $(k = 1, \ldots, K)$, has asymptotic variance smaller than or equal to that of any $\phi_k\{\hat{\beta}(\eta, \hat{\alpha})\}$ with η fixed but arbitrary, if the missingness model is correct and even if the outcome regression model is incorrect;

Property 8. they have asymptotic variance smaller than or equal to that of $\hat{\beta}_{\rm ipw}$ if the missingness model is correct even if the outcome regression model is incorrect.

For β the population mean of a scalar outcome, as in Tan (2008, 2010a) and Cao et al. (2009), or for any other scalar parameter, Property 7 with $\phi_1(\beta) = \beta$ is the same as Property 3. For β of dimension 2 or more, Property 7 ensures enhanced efficiency for estimation of a finite set of target scalar features of the vector β specified by the data analyst. For example, if in a two-group comparison analysis, $h(Z; \beta) = \text{expit}(\beta_1 + \beta_2 Z)$ with Z binary, one would choose $\phi_1(\beta) = h(0; \beta)$, $\phi_2(\beta) = h(1; \beta)$ and $\phi_3(\beta) = h(1; \beta) - h(0; \beta)$ if one wants to ensure

enhanced efficient estimation of the group means and of their difference. A more ambitious goal would be to construct estimators that satisfy Property 7 for all smooth scalar functions ϕ, not just for a finite number of them, or equivalently, Property 3 even if the dimension of β is 2 or more. Whether or not this can be accomplished remains an open question.

For estimating regression models with missing data, several procedures exist that satisfy some, but not all of Properties 4–8. The standard double robust estimator $\hat{\beta}(\hat{\eta}, \hat{\alpha})$ satisfies Properties 4 and 5 but not 6–8. Bang & Robins (2005) estimator and estimators derived from application of targeted maximum likelihood methodology (van der Laan & Rubin, 2006; van der Laan, 2010) satisfy Properties 4–6 but not 7 and 8. The one step corrected estimator in § 2.5 of van der Laan & Robins (2003) satisfies Properties 5 and 8 but not 4, 6 and 7. The estimators in Tan (2010b) satisfy Properties 4, 5 and 8 but not 6 and 7.

2·2. *Proposed estimator*

To compute $\hat{\beta}_{\mathrm{dr}}$ it is required that the parameter η of model (3) has dimension q greater than or equal to the dimension p of β_0. If this is not the case, one first augments model (3) by including additional covariates based on transformations of L, e.g., powers of L. The computation of $\hat{\beta}_{\mathrm{dr}}$ is carried out in the following three steps:

Step 1. Estimate η by $\hat{\eta}_{\mathrm{or}}$ solving the p equations

$$E_n[\omega(\hat{\alpha})b(Z)\{Y - m(L; \eta)\}] = 0, \tag{10}$$

and the additional $q - p$ equations

$$E_n[Ad(L; \eta)\{Y - m(L; \eta)\}] = 0, \tag{11}$$

where $d(L; \eta)$ is any user-specified, $(q - p) \times 1$ function, say $d(L; \eta) = (\partial m(L; \eta)/\partial\eta_1, \ldots, \partial m(L; \eta)/\partial\eta_{q-p})^{\mathrm{T}}$. Compute $\hat{\beta}_{\mathrm{or}} = \hat{\beta}_{\mathrm{reg}}(\hat{\eta}_{\mathrm{or}})$ solving the outcome regression equation (4) with $\eta = \hat{\eta}_{\mathrm{or}}$.

Step 2. For each k $(k = 1, \ldots, K)$, compute

$$\tilde{\eta}_k = \arg\min_{\eta} E_n([\hat{I}_k\{M(\hat{\alpha}, \hat{\beta}_{\mathrm{or}}) - U(\eta, \hat{\alpha}, \hat{\beta}_{\mathrm{or}}) - \hat{\rho}_{\eta}^{\mathrm{T}} S(\hat{\alpha})\}]^2), \tag{12}$$

where

$$\begin{aligned} S(\alpha) &= \partial \log f(A \mid L; \alpha)/\partial\alpha, \\ \hat{I}_k &= \partial\phi_k(\beta)/\partial\beta^{\mathrm{T}}|_{\hat{\beta}_{\mathrm{or}}} E_n\{\partial M(\hat{\alpha}, \beta)/\partial\beta^{\mathrm{T}}|_{\hat{\beta}_{\mathrm{or}}}\}^{-1}, \\ M(\alpha, \beta) &= \omega(\alpha)b(Z)\{Y - h(Z; \beta)\}, \\ U(\eta, \alpha, \beta) &= g^b_{\eta,\alpha,\beta}(A, L) - E_{\alpha}\{g^b_{\eta,\alpha,\beta}(A, L) \mid L\} \end{aligned}$$

and

$$\hat{\rho}_{\eta}^{\mathrm{T}} = E_n[\{M(\hat{\alpha}, \hat{\beta}_{\mathrm{or}}) - U(\eta, \hat{\alpha}, \hat{\beta}_{\mathrm{or}})\}S(\hat{\alpha})^{\mathrm{T}}]E_n[S(\hat{\alpha})S(\hat{\alpha})^{\mathrm{T}}]^{-1}.$$

Step 3. Compute the maximum likelihood estimator $(\tilde{\alpha}, \tilde{\delta})$ of (α, δ) in the augmented missingness model

$$f(A \mid L; \alpha, \delta) = c(L; \alpha, \delta) f(A \mid L; \alpha) \exp\left\{ \sum_{k=1}^{K+1} \delta_k^{\mathrm{T}} u_k(A, L) \right\}, \tag{13}$$

where $u_k(A, L) = U(\tilde{\eta}_k, \hat{\alpha}, \hat{\beta}_{\mathrm{or}})$ $(k = 1, \ldots, K)$, $u_{K+1}(A, L) = U(\hat{\eta}_{\mathrm{or}}, \hat{\alpha}, \hat{\beta}_{\mathrm{or}})$ and $c(L; \alpha, \delta)$ is the normalizing constant. Estimate η by $\tilde{\eta}_{\mathrm{or}}$ jointly solving (11) and (10) with $\omega(\hat{\alpha})$ replaced with $\omega(\tilde{\alpha}, \tilde{\delta}) = A f(A \mid L; \tilde{\alpha}, \tilde{\delta})^{-1}$. Finally, $\hat{\beta}_{\mathrm{dr}}$ is the estimator $\hat{\beta}_{\mathrm{reg}}(\tilde{\eta}_{\mathrm{or}})$ solving the outcome regression equation (4) with $\eta = \tilde{\eta}_{\mathrm{or}}$.

The estimator $\hat{\beta}_{\mathrm{or}}$ of β returned by Step 1 is the extension to the regression setting of the so-called weighted least squares outcome regression double-robust estimator of a population mean of Kang & Schafer (2007, Equation (10)). Step 2 follows Rubin & van der Laan's (2008) prescription to compute, separately for each k, an estimator $\tilde{\eta}_k$ of η targeted at minimizing the asymptotic variance of $\phi_k\{\hat{\beta}(\eta, \hat{\alpha})\}$ if model (5) is correct. Under this assumption, the empirical mean in (12) is a consistent estimator of the asymptotic variance of $\phi_k\{\hat{\beta}(\eta, \hat{\alpha})\}$. Step 3 simply repeats Step 1, after re-estimating the missingness probabilities under the extended model (13). The subscripts in $\tilde{\eta}_{\mathrm{or}}$ and $\hat{\eta}_{\mathrm{or}}$ are a reminder that when either is replaced for η in $\hat{\beta}(\eta; \hat{\alpha})$, it yields estimators solving outcome regression estimating equations.

The vector $\hat{\theta} = \{\hat{\alpha}, \hat{\beta}_{\mathrm{or}}, \hat{\eta}_{\mathrm{or}}, \tilde{\eta}_{\mathrm{or}}, \hat{\beta}_{\mathrm{dr}}, \tilde{\alpha}, \tilde{\delta}, (\tilde{\eta}_k, \hat{\rho}_{\tilde{\eta}_k})_{1 \leqslant k \leqslant K}\}$ solves a system of estimating equations $E_n\{\Psi(X; \theta)\} = 0$. Under the conditions of van der Vaart (2000, Theorems 5.9 and 5.21), $\hat{\theta} - \theta^* = -V_{\theta^*}^{-1} E_n\{\Psi(X; \theta^*)\} + o_p(n^{-1/2})$, for θ^* satisfying $E\{\Psi(X; \theta^*)\} = 0$ and $V_\theta = \partial E\{\Psi(X; \theta)\}/\partial\theta$. In what follows, we will assume that these conditions hold and argue that the estimator $\hat{\beta}_{\mathrm{dr}}$ satisfies the Properties 4–8.

For Property 4, the estimator $\tilde{\eta}_{\mathrm{or}}$ converges in probability to a solution of the joint equations $E[\omega(\alpha^*, \delta^*) b(Z)\{Y - m(L; \eta)\}] = 0$ and $E[A d(L; \eta)\{Y - m(L; \eta)\}] = 0$ where $(\alpha^*, \delta^*) = \mathrm{plim}\,(\tilde{\alpha}, \tilde{\delta})$. If model (3) is correct, η_0 is a solution to this system. Thus $\hat{\beta}_{\mathrm{dr}}$, being equal to the estimator $\hat{\beta}_{\mathrm{reg}}(\tilde{\eta}_{\mathrm{or}})$, is consistent for β_0 and asymptotically normal if model (3) is correct and the preceding joint population equations have a unique solution. On the other hand, $\hat{\beta}_{\mathrm{dr}}$ is equal to $\hat{\beta}\{\tilde{\eta}_{\mathrm{or}}, (\tilde{\alpha}, \tilde{\delta})\}$ solving (7) with $\eta = \tilde{\eta}_{\mathrm{or}}$ and $\alpha = (\tilde{\alpha}, \tilde{\delta})$ because (7) is the same as equations

$$E_n[b(Z)\{m(L; \eta) - h(Z; \beta)\}] + E_n[\omega(\alpha) b(Z)\{Y - m(L; \eta)\}] = 0, \tag{14}$$

and, by construction of $\hat{\beta}_{\mathrm{dr}}$ and $\tilde{\eta}_{\mathrm{or}}$, $E_n[b(Z)\{m(L; \tilde{\eta}_{\mathrm{or}}) - h(Z; \hat{\beta}_{\mathrm{dr}})\}] = 0$ and $E_n[\omega(\tilde{\alpha}, \tilde{\delta}) b(Z)\{Y - m(L; \tilde{\eta}_{\mathrm{or}})\}] = 0$. When model (5) is correct, model (13) is also correct with a true value of δ equal to 0. Consequently, when (5) is correct $\hat{\beta}\{\tilde{\eta}_{\mathrm{or}}, (\tilde{\alpha}, \tilde{\delta})\}$, just as any estimator solving (7), is consistent for β_0 and asymptotically normal.

Property 5 holds because, as indicated above, $\hat{\beta}_{\mathrm{dr}}$ solves equation (7) with η and α replaced by estimators that converge to the true parameter values when the models they index are correct.

Property 6 holds by construction.

For Property 7, under model (5), $\hat{\beta}(\eta, \hat{\alpha})$ satisfies the expansion

$$\hat{\beta}(\eta, \hat{\alpha}) - (\beta_0) = E_n[I\{M_0 - U_0(\eta) - \rho_\eta^{\mathrm{T}} S(\alpha_0)\}] + o_p(n^{-1/2}),$$

where $M_0 = M(\alpha_0, \beta_0)$, $U_0(\eta) = U(\eta, \alpha_0, \beta_0)$, $I = E\{\partial M(\alpha_0, \beta)/\partial\beta^{\mathrm{T}}|_{\beta_0}\}^{-1}$ and

$$\rho_\eta^{\mathrm{T}} = E[\{M_0 - U_0(\eta)\} S^{\mathrm{T}}(\alpha_0)] E\{S(\alpha_0) S^{\mathrm{T}}(\alpha_0)\}^{-1}$$

is the population least squares coefficient in the multivariate regression of $M_0 - U_0(\eta)$ on the components of the score vector $S(\alpha_0)$ (Robins et al., 1994). Thus, an application of the delta method yields that

$$\text{avar}[\phi_k\{\hat{\beta}(\eta, \hat{\alpha})\}] = \sigma_k^2(\eta) = \min_{\rho} E[[I_k\{M_0 - U_0(\eta) - \rho^{\mathrm{T}} S(\alpha_0)\}]^2], \tag{15}$$

where $I_k = \partial\phi_k(\beta)/\partial\beta^{\mathrm{T}}|_{\beta_0} E\{\partial M(\alpha_0, \beta)/\partial\beta^{\mathrm{T}}|_{\beta_0}\}^{-1}$, and in the sequel avar$(\cdot)$ stands for the variance of the limiting distribution of $\cdot$. As the dimension of the vector α indexing the missingness model increases, so does the dimension of $S(\alpha_0)$ and consequently $\sigma_k^2(\eta)$ decreases. With this in mind, we construct in Step 3 an augmented missingness model choosing to enlarge model (5) with just the right additional covariates so as to ensure that the resulting estimator of β_0 is at least as efficient asymptotically as $\phi_k\{\hat{\beta}(\eta_k^*, \hat{\alpha})\}$ $(k = 1, \ldots, K)$, where

$$\eta_k^* = \arg\min_{\eta} \sigma_k^2(\eta).$$

Specifically, when model (5) is correct, so too is the enlarged model (13) with true parameter values α_0 and $\delta_{0,k} = 0$ $(k = 1, \ldots, K+1)$. It then follows from (8) and the fact that $\hat{\beta}_{\text{dr}} = \hat{\beta}\{\tilde{\eta}_{\text{or}}, (\tilde{\alpha}, \tilde{\delta})\}$ that

$$\hat{\beta}_{\text{dr}} - \beta_0 = E_n\left[I\left\{M_0 - U_0(\eta_{\text{or}}^{**}) - \rho^{*\mathrm{T}} S_\alpha^*(\alpha_0, \delta_0) - \sum_{k=1}^{K+1} \nu_k^{*\mathrm{T}} S_{\delta_k}^*(\alpha_0, \delta_0)\right\}\right] + o_p(n^{-1/2}), \tag{16}$$

where $\eta_{\text{or}}^{**} = \text{plim}\tilde{\eta}_{\text{or}}$, $S_\alpha^*(\alpha, \delta) = \partial \log f^*(A \mid L; \alpha, \delta)/\partial\alpha$ and $S_{\delta_k}^*(\alpha, \delta) = \partial \log f^*(A \mid L; \alpha, \delta)/\partial\delta_k$ are the scores in the model

$$f^*(A \mid L; \alpha, \delta) = c^*(L; \alpha, \delta) f(A \mid L; \alpha) \exp\left\{\sum_{k=1}^{K+1} \delta_k^{\mathrm{T}} U_0(\eta_k^*) + \delta_{K+1}^{\mathrm{T}} U_0(\eta_{\text{or}}^*)\right\}, \tag{17}$$

with $c^*(L; \alpha, \delta)$ the normalizing constant and $\eta_{\text{or}}^* = \text{plim}\hat{\eta}_{\text{or}}$. Furthermore, $(\rho^{*\mathrm{T}}, \nu_1^{*\mathrm{T}}, \ldots, \nu_{K+1}^{*\mathrm{T}})$ is the population least squares constant in the regression of $M_0 - U_0(\eta_{\text{or}}^{**})$ on $S_\alpha^*(\alpha_0, \delta_0), S_{\delta_1}^*(\alpha_0, \delta_0), \ldots, S_{\delta_{K+1}}^*(\alpha_0, \delta_0)$. Now, because of the precise form of model (17), it holds that $S_\alpha^*(\alpha_0, \delta_0) = S(\alpha_0)$, $S_{\delta_k}^*(\alpha_0, \delta_0) = U_0(\eta_k^*)$ $(k = 1, \ldots, K)$, and $S_{\delta_{K+1}}^*(\alpha_0, \delta_0) = U_0(\eta_{\text{or}}^*)$. Furthermore, $U_0(\eta_{\text{or}}^{**}) = U_0(\eta_{\text{or}}^*)$ because, as argued in the Appendix, when model (5) holds, $\eta_{\text{or}}^* = \eta_{\text{or}}^{**}$. Thus, expansion (16) reduces to

$$\hat{\beta}_{\text{dr}} - \beta_0 = E_n\left[I\left\{M_0 - \rho^{*\mathrm{T}} S(\alpha_0) - \sum_{k=1}^{K} \nu_k^{*\mathrm{T}} U_0(\eta_k^*) - \nu_{K+1}^{*\mathrm{T}} U_0(\eta_{\text{or}}^*)\right\}\right] + o_p(n^{-1/2}), \tag{18}$$

with $(\rho^{*\mathrm{T}}, \nu_1^{*\mathrm{T}}, \ldots, \nu_{K+1}^{*\mathrm{T}})$ re-defined as the population least squares coefficient in the regression of M_0 on $S(\alpha_0), U_0(\eta_1^*), \ldots, U_0(\eta_K^*)$ and $U_0(\eta_{\text{or}}^*)$.

An application of the delta method then yields

$$\text{avar}\{\phi_k(\hat{\beta}_{\text{dr}})\} = \min_{(\rho, \nu)} E\left(\left[I_k\left\{M_0 - \rho^{\mathrm{T}} S(\alpha_0) - \sum_{k=1}^{K} \nu_k^{\mathrm{T}} U_0(\eta_k^*) - \nu_{K+1}^{\mathrm{T}} U_0(\eta_{\text{or}}^*)\right\}\right]^2\right). \tag{19}$$

This shows that $\text{avar}\{\phi_k(\hat{\beta}_{\text{dr}})\} \leqslant \text{avar}[\phi_k\{\hat{\beta}(\eta, \hat{\alpha})\}]$ for any η because by definition of η_k^*, the smallest possible $\text{avar}[\phi_k\{\hat{\beta}(\eta, \hat{\alpha})\}]$ is the right-hand side of (15) evaluated at η_k^*, and this quantity is greater than or equal to the right-hand side of the last display because this is a minimum over a larger set.

Property 8 follows by noticing that

$$\hat{\beta}_{\text{IPW}} - \beta_0 = E_n[I\{M_0 - \rho^{**\text{T}} S(\alpha_0)\}] + o_p(n^{-1/2}),$$

where $M_0 - \rho^{**\text{T}} S(\alpha_0)$ is the residual from the population regression of M_0 on $S(\alpha_0)$. This residual has variance larger than or equal, in the positive definite sense, the residual between curly brackets in (18) as the latter is the residual from the regression of M_0 on covariates that include $S(\alpha_0)$.

The following remarks help clarify our construction. First, Step 1 is needed in order to include the covariate $u_{K+1}(A, L)$ in model (13). Without this covariate, the asymptotic variance of $\phi_k(\hat{\beta}_{\text{dr}})$ would be equal to the variance of the expression between squared brackets in (16) with I_k instead of I and without the term $\nu_{K+1}^{\text{T}} S^*_{\delta_{K+1}}(\alpha_0, \delta_0)$. But this variance would not necessarily be smaller than the right-hand side of (15) evaluated at η_k^*. Second, we can modify Step 3 to yield $\hat{\beta}_{\text{dr}}$ additionally as efficient as $\hat{\beta}_{g,\text{aipw}}$ for any specified g by simply adding the term $\delta_{K+2} u_{K+2}(A, L)$, where $u_{K+2}(A, L) = [g(A, L) - E_{\hat{\alpha}}\{g(A, L) \mid L\}]$, to the exponential tilt in model (13). Third, the computation of $\tilde{\eta}_{\text{or}}$ in Step 3 depends on $(\tilde{\alpha}, \tilde{\delta})$ only through the $f(A_i \mid L_i; \tilde{\alpha}, \tilde{\delta})$ $(i = 1, \ldots, n)$. If the outcome regression model (3) is correctly specified, then some or all of the $u_k(A, L)$ $(k = 1, \ldots, K + 1)$, may converge in probability to the same function of (L, A) and thus they may be highly collinear. In such a case, $\tilde{\delta}$ may not be unique. However, $\sum_{k=1}^{K+1} \tilde{\delta}_k u_k(A, L)$, and hence $f(A \mid L; \tilde{\alpha}, \tilde{\delta})$, will still be unique. Formula (19) for the asymptotic variance of $\phi_k(\hat{\beta}_{\text{dr}})$ remains valid with the provision that some or all of the $U_0(\eta_k^*)$ $(k = 1, \ldots, K)$, and $U_0(\eta_{\text{or}}^*)$ may be equal. This provision does not invalidate the preceding arguments justifying that $\hat{\beta}_{\text{dr}}$ satisfies Properties 7 and 8.

Standard empirical sandwich variance techniques could in principle be used to derive an estimator that is consistent for the asymptotic variance of $\hat{\beta}_{\text{dr}}$ regardless of the validity of models (5) or (3), because, as indicated earlier, $\hat{\theta} = \{\hat{\alpha}, \hat{\beta}_{\text{or}}, \hat{\eta}_{\text{or}}, \tilde{\eta}_{\text{or}}, \hat{\beta}_{\text{dr}}, \tilde{\alpha}, \tilde{\delta}, (\tilde{\eta}_k, \hat{\rho}_{\tilde{\eta}_k})_{1 \leqslant k \leqslant K}\}$ solves an estimating equation $E_n\{\Psi(\theta)\} = 0$. However, finding the analytic expression for Ψ would be cumbersome. Nevertheless, the nonparametric bootstrap can be used to compute a consistent variance estimator of $\hat{\beta}_{\text{dr}}$ because $\hat{\theta}$ is regular and asymptotically linear (Gill, 1989).

Example 1. Consider estimation of $\beta_0 = E(Y)$ with Y binary. In this case, $h(\beta) = \beta$ and Z is absent in model (1). Suppose we assume that (5) and (3) are logistic regressions with covariates L and intercept. The score for α is $S(\alpha) = \tilde{L}\,\text{expit}(\alpha^{\text{T}}\tilde{L})\{\omega(\alpha) - 1\}$. The function $b(\cdot)$ in equation (4) is a scalar constant since Z is absent. All bs yield the same estimator of β_0, so we assume $b = 1$. In Step 1, $\hat{\eta}_{\text{or}}$ solves (10) and $E_n[A\tilde{L}_r\{Y - \text{expit}(\eta^{\text{T}}\tilde{L})\}] = 0$ $(r = 1, \ldots, q - 1)$ yielding $\hat{\beta}_{\text{or}} = E_n\{\text{expit}(\hat{\eta}_{\text{or}}^{\text{T}}\tilde{L})\}$. In Step 2, $K = 1$, $\phi_1(\beta) = \beta$ and $I_1 = 1$. Furthermore, $U(\eta, \alpha, \beta) = \{\omega(\alpha) - 1\}\{\text{expit}(\eta^{\text{T}}\tilde{L}) - \beta\}$. Consequently,

$$\tilde{\eta}_1 = \arg\min_{\eta} E_n([\omega(\hat{\alpha})Y - \{\omega(\hat{\alpha}) - 1\}\{\text{expit}(\eta^{\text{T}}\tilde{L}) - \hat{\rho}_{\eta}^{\text{T}}\tilde{L}\,\text{expit}(\hat{\alpha}^{\text{T}}\tilde{L})\} + \hat{\beta}_{\text{or}}]^2).$$

Model (13) of Step 3 is a logistic regression with intercept and covariates L, $x_1(L) = \text{expit}(\hat{\alpha}^{\text{T}}\tilde{L})^{-1}\{\text{expit}(\tilde{\eta}_1^{\text{T}}\tilde{L}) - \hat{\beta}_{\text{or}}\}$ and $x_2(L) = \text{expit}(\hat{\alpha}^{\text{T}}\tilde{L})^{-1}\{\text{expit}(\hat{\eta}_{\text{or}}^{\text{T}}\tilde{L}) - \hat{\beta}_{\text{or}}\}$. The estimator $\tilde{\eta}_{\text{or}}$ is computed just like $\hat{\eta}_{\text{or}}$, except that in equation (10) $\omega(\hat{\alpha})$ is replaced by $\omega(\tilde{\alpha}, \tilde{\delta}) = A\,\text{expit}\{\tilde{\alpha}^{\text{T}}\tilde{L} + \tilde{\delta}_1 x_1(L) + \tilde{\delta}_2 x_2(L)\}^{-1}$. Finally, $\hat{\beta}_{\text{dr}} = E_n\{\text{expit}(\tilde{\eta}_{\text{or}}^{\text{T}}\tilde{L})\}$, which has asymptotic

variance under model (5),

$$\text{avar}(\hat{\beta}_{\text{dr}}) = \min_{(\rho,\nu)} E[\{M_0 - \rho^{\mathrm{T}} S(\alpha_0) - \nu_1^{\mathrm{T}} U_0(\eta_1^*) - \nu_2^{\mathrm{T}} U_0(\eta_{\text{or}}^*)\}^2],$$

with $M_0 = \omega(\alpha_0)(Y - \beta_0)$, $U_0(\eta_j^*) = \{\omega(\alpha_0) - 1\}\{\text{expit}(\eta_j^{*\mathrm{T}} \tilde{L}) - \beta_0\}$ and $\eta_j^* = \text{plim}\, \tilde{\eta}_j$ for $j = 1$, or. Interestingly, the estimator of Cao et al. (2009) has asymptotic variance $\min_\rho E[\{M_0 - \rho^{\mathrm{T}} S(\alpha_0) - U_0(\eta_1^*)\}^2]$, which is generally strictly larger than $\text{avar}(\hat{\beta}_{\text{dr}})$ due to the nonlinear dependence of $U_0(\eta_1^*)$ on η_1^*.

3. CAUSAL INFERENCE IN POINT-EXPOSURE STUDIES

We now consider estimation of marginal structural mean models for causal inference in point exposure observational studies. Suppose that in an observational study with n subjects drawn at random from a population of interest, we observe $(A_i, Y_i, L_i^{\mathrm{T}})$ $(i = 1, \ldots, n)$, independent and identically distributed across i, where Y_i is a scalar outcome, A_i is a treatment variable taking values in a set $\mathcal{A}$ and L_i is a vector of pre-treatment confounding variables. For each $a \in \mathcal{A}$, let $Y_{(a),i}$ be the potential outcome of the ith unit under treatment a. We make the usual consistency assumption that $Y_{(A_i),i} = Y_i$. Suppose that Z_i is a subvector of L_i, possibly empty. A marginal structural mean model postulates that

$$E(Y_{(a)} \mid Z) = h(a, Z; \beta_0),$$

where $h(\cdot;\cdot)$ is a known smooth function and β_0 is an unknown $p \times 1$ parameter vector (Robins, 1999). For example, $h(a, Z; \beta) = \beta_1 + \beta_2 a + \beta_3 a Z$. We examine estimation of β_0 from data $(A_i, Y_i, L_i^{\mathrm{T}})$ $(i = 1, \ldots, n)$, under the no-unmeasured confounders assumption which states that $Y_{(a)}$ and A are conditionally independent given L for all $a \in \mathcal{A}$. Under this assumption, regarding $Y_{(a)}$ as an outcome variable that is observed and equal to Y only in units that received treatment A equal to a it holds, like in § 2·1, that $E(Y_{(a)} \mid Z) = E\{E(Y \mid L, A = a) \mid Z\} = E(\omega_a Y \mid Z)$ where $\omega_a = I_a(A) f(A \mid L)^{-1}$ and $I_a(A) = 1$ if $A = a$ and $I_a(A) = 0$ otherwise.

In this setting, an outcome regression estimator $\hat{\beta}_{\text{reg}}(\hat{\eta})$ is the solution of

$$\int_{\mathcal{A}} E_n[b(a, Z)\{m(a, L; \hat{\eta}) - h(a, Z; \beta)\}]\, \mathrm{d}\mu(a) = 0, \tag{20}$$

where $b(a, z)$ is a $p \times 1$ user-specified vector function, μ denotes the Lebesgue measure when A is continuous and the counting measure if A is discrete, and $\hat{\eta}$ is an estimator of η_0 under a postulated outcome regression model

$$E(Y \mid L, A) = m(A, L; \eta_0), \tag{21}$$

which simultaneously parameterizes the separate outcome regressions $E(Y \mid L, A = a)$ for all $a \in \mathcal{A}$. If model (21) is correct, then $\hat{\beta}_{\text{reg}}(\hat{\eta})$ solving (20) is consistent for β_0 and asymptotically normal provided $\hat{\eta}$ is consistent for η_0 and asymptotically normal. Alternatively, positing a model (5) for the treatment mechanism we can consider estimating equations (6) and (7) where $b(Z)$, $m(L; \eta)$, $h(Z; \beta)$ are replaced with $b(A, Z)$, $m(A, L; \eta)$, $h(A, Z; \beta)$, $\hat{\eta}$ is now an estimator that is consistent for η_0 when model (21) is correct and with the re-definitions

$$\omega = \omega_A = f(A \mid L)^{-1}, \quad \omega(\alpha) = f(A \mid L; \alpha)^{-1}.$$

With these modifications, we obtain the estimators $\hat{\alpha}$, $\hat{\beta}_{g,\mathrm{aipw}}$, $\hat{\beta}_{\mathrm{ipw}}$, $\hat{\beta}(\eta, \alpha)$ and $\hat{\beta}(\hat{\eta}, \alpha)$ as in § 2·1. Robins (1999) showed that (8) holds and Scharfstein et al. (1999) showed that $\hat{\beta}(\hat{\eta}, \hat{\alpha})$ satisfies Properties 4 and 5 of § 2·1 where the words missingness model are replaced by treatment model and the outcome regression model is (21) rather than (3).

As in the missing data case, standard double-robust estimators $\hat{\beta}(\hat{\eta}, \hat{\alpha})$ with $\hat{\eta}$ the ordinary or iteratively reweighted least squares estimator of η_0 based on all sampled units may not exist, because equation (7) evaluated at $\hat{\eta}$ and $\hat{\alpha}$ may not have a solution. Furthermore, $\hat{\beta}(\hat{\eta}, \hat{\alpha})$ does not have the advantageous efficiency Properties 7 and 8 of § 2·2. In the present setting, we can define the estimator $\hat{\beta}_{\mathrm{dr}}$ identically as in § 2·2, but with the re-definitions and replacements indicated in the preceding paragraph. Arguments identical to those given in § 2·2 imply that $\hat{\beta}_{\mathrm{dr}}$ satisfies Properties 4–8 of § 2·2 with the outcome regression equation referred to in Property 6 being now (20).

4. Monotone missing data in longitudinal studies

We now turn to double-robust estimation in regression models for longitudinal data with drop-out. The intended data are n independent and identically distributed copies of $\bar{L}_J = (L_0^{\mathrm{T}}, L_1^{\mathrm{T}}, \ldots, L_J^{\mathrm{T}})^{\mathrm{T}}$ where $\bar{B}_j$ denotes $(B_0, \ldots, B_j)$ throughout. The vector L_j denotes the data we intend to measure at the jth occasion on a sample unit. Let C denote the drop-out occasion: $C = j$ on a given sample unit if and only if we observe $\bar{L}_j$ for that unit. We assume L_0 is always observed, so $C > 0$. The outcome of interest Y is $r(\bar{L}_J)$, where $r(\cdot)$ is a user-specified function which for simplicity we assume is scalar valued; for example, $r(\bar{L}_J)$ is some component of the vector L_J. The goal is to estimate the $p \times 1$ parameter vector β_0 of a regression model (4) where Z is a subvector of L_0, possibly empty, from n independent and identically distributed copies of $(C, \bar{L}_C^{\mathrm{T}})$ under the missing at random assumption

$$f(A_j \mid \bar{A}_{j-1}, \bar{L}_J) = f(A_j \mid \bar{A}_{j-1}, \bar{L}_{j-1}) \quad (j = 1, \ldots, J),$$

where A_j is the on study indicator, i.e., $A_j = 1$ if $C \geqslant j$ and $A_j = 0$ otherwise. Provided $\mathrm{pr}(A_j = 1 \mid A_{j-1} = 1, \bar{L}_{j-1}) > 0$ $(j = 1, \ldots, J)$, $E(Y \mid Z)$ can be expressed in two different ways (Bang & Robins, 2005), each leading to a different estimation strategy,

$$E(Y \mid Z) = E(H_0 \mid Z) = E(\omega_J Y | Z),$$

with H_0 defined from the recursion $H_J = Y$, and $H_{j-1} = E(H_j \mid A_j = 1, \bar{L}_{j-1})$ $(j = J, \ldots, 1)$, and

$$\omega_j = A_j \times \{f(A_1 \mid \bar{A}_0, \bar{L}_0) \times \cdots \times f(A_j \mid \bar{A}_{j-1}, \bar{L}_{j-1})\}^{-1} \quad (j = 0, \ldots, J-1).$$

To generalize $\hat{\beta}_{\mathrm{reg}}(\hat{\eta})$ to the longitudinal setting, we posit outcome regression models

$$E(H_j \mid A_j = 1, \bar{L}_{j-1}) = m_j(\bar{L}_{j-1}; \eta_0) \quad (j = 1, \ldots, J), \tag{22}$$

with η_0 an unknown $q \times 1$ parameter vector and $m_j(\cdot; \cdot)$ known functions, for instance $m_j(\bar{L}_{j-1}; \eta) = \Phi^{-1}\{\eta^{\mathrm{T}} s_j(\bar{L}_{j-1})\}$ for some link function $\Phi(\cdot)$ and some user-specified functions

$s_j(\cdot)$, and we estimate η_0 by some $\hat{\eta}$, for example, by $\hat{\eta}$ solving

$$E_n\left[A_J d_J(\bar{L}_J;\eta)\{Y - m_J(\bar{L}_J;\eta)\} + \sum_{j=1}^{J-1} A_j d_j(\bar{L}_j;\eta)\{m_{j+1}(\bar{L}_{j+1};\eta) - m_j(\bar{L}_j;\eta)\}\right] = 0, \tag{23}$$

where $d_j(\bar{L}_j;\eta) = \partial m_j(\bar{L}_j;\eta)/\partial\eta$ and $\sum_{j=1}^{J-1}(\cdot) = 0$ if $J = 1$. The outcome regression estimator $\hat{\beta}_{\mathrm{reg}}(\hat{\eta})$ then solves (4) with $m(L;\eta)$ re-defined as $m_1(\bar{L}_0;\eta)$ here and throughout this section. If (22) holds, $\hat{\beta}_{\mathrm{reg}}(\hat{\eta}) - \beta_0 = O_p(n^{-1/2})$ provided $\hat{\eta}$ solves (23) or, more generally, provided $\hat{\eta} - \eta_0 = O_p(n^{-1/2})$.

Alternatively, to generalize $\hat{\beta}_{g,\mathrm{aipw}}$ to the longitudinal setting, we posit drop-out models

$$f(A_j \mid \bar{A}_{j-1}, \bar{L}_{j-1}) = A_{j-1} f(A_j \mid A_{j-1} = 1, \bar{L}_{j-1};\alpha_0) \quad (j = 1,\ldots,J) \tag{24}$$

for instance, $f(A_j \mid A_{j-1} = 1, \bar{L}_{j-1};\alpha) = \mathrm{expit}\{\alpha^{\mathrm{T}} t_j(\bar{L}_{j-1})\}$ for some user-specified functions $t_j(\cdot)$, and we compute the maximum likelihood estimator $\hat{\alpha}$ of α_0. Then, we compute $\hat{\beta}_{g,\mathrm{aipw}}$ solving, with α evaluated at $\hat{\alpha}$, the equation

$$E_n[\omega_J(\alpha)b(Z)\{Y - h(Z;\beta)\}] - E_n \times \left[\sum_{j=1}^{J} g_j(\bar{A}_j, \bar{L}_{j-1}) - E_\alpha\{g_j(\bar{A}_j, \bar{L}_{j-1}) \mid \bar{A}_{j-1}, \bar{L}_{j-1}\}\right] = 0$$

for some $p \times 1$ vector functions $g_1(\cdot),\ldots,g_J(\cdot)$ and $b(\cdot)$. Under regularity conditions, $\hat{\beta}_{g,\mathrm{aipw}}$ is consistent for β_0 and asymptotically normal (Robins & Rotnitzky, 1995) when model (24) holds with asymptotic variance no smaller than that of $\hat{\beta}_{g_{\mathrm{opt}},\mathrm{aipw}}$ where $g^b_{\mathrm{opt,j}}(\bar{A}_j, \bar{L}_{j-1}) = \omega_j b(Z)\{E(H_j \mid A_j = 1, \bar{L}_{j-1}) - h(Z;\beta_0)\}$ (Robins et al., 1994). The extension of $\hat{\beta}(\eta,\alpha)$ to the longitudinal setting solves

$$E_n[\omega_J(\alpha)b(Z)\{Y - h(Z;\beta)\}] - E_n\left\{\sum_{j=1}^{J} G^b_{\eta,\alpha,\beta,j} - E_\alpha(G^b_{\eta,\alpha,\beta,j} \mid \bar{A}_{j-1}, \bar{L}_{j-1})\right\} = 0, \tag{25}$$

with

$$G^b_{\eta,\alpha,\beta,j} \equiv g^b_{\eta,\alpha,\beta,j}(\bar{A}_j, \bar{L}_{j-1}) \equiv \omega_j(\alpha)b(Z)\{m_j(\bar{L}_{j-1};\eta) - h(Z;\beta)\},$$

and $\omega_j(\alpha) = A_j \times \{f(A_1 \mid \bar{A}_0, \bar{L}_0;\alpha) \times \cdots \times f(A_j \mid \bar{A}_{j-1}, \bar{L}_{j-1};\alpha)\}^{-1}$. The estimator $\hat{\beta}(\hat{\eta},\hat{\alpha})$ satisfies (8) and Properties 4 and 5 with models (24) and (22) instead of (5) and (3) (Robins, 2000). However, as in § 2, for most estimators $\hat{\eta}$, e.g., for $\hat{\eta}$ solving (23), equation (25) evaluated at $\hat{\eta}$ and $\hat{\alpha}$ may not have a solution. Furthermore, $\hat{\beta}(\hat{\eta},\hat{\alpha})$ does not satisfy the efficiency Properties 7 and 8.

An estimator $\hat{\beta}_{\mathrm{dr}}$ satisfying Properties 4–8 of § 2·1, with models (24) and (22) instead of (5) and (3) and Property 8 can be constructed implementing the following generalization of the three step procedure of § 2·2 to the longitudinal setting.

Step 1. Compute $\hat{\beta}_{\mathrm{or}} = \hat{\beta}_{\mathrm{reg}}(\hat{\eta}_{\mathrm{or}})$ solving (4) in which $m(L;\eta)$ re-defined as $m_1(\bar{L}_0;\eta)$ and with $\eta = \hat{\eta}_{\mathrm{or}}$, the estimator of the $q \times 1$ parameter vector η of model (22) ($q \geqslant p$) solving the p

equations

$$E_n\left(b(Z)\left[\sum_{j=1}^{J-1}\omega_j(\hat\alpha)\{m_{j+1}(\bar L_j;\eta)-m_j(\bar L_{j-1};\eta)\}+\omega_J(\hat\alpha)\{Y-m_J(\bar L_J,\eta)\}\right]\right)=0,$$

and the first $q-p$ equations in the system (23).

Step 2. For each k compute

$$\tilde\eta_k=\arg\min_\eta E_n([\hat I_k\{M(\hat\alpha,\hat\beta_{\rm or})-U(\eta,\hat\alpha,\hat\beta_{\rm or})-\hat\rho_\eta^{\rm T}S(\hat\alpha)\}]^2),$$

where $U(\eta,\alpha,\beta)=\sum_{j=1}^J U_j(\eta,\alpha,\beta)$ with

$$U_j(\eta,\alpha,\beta)=g^b_{\eta,\alpha,\beta,j}(\bar A_j,\bar L_{j-1})-E_\alpha\{g^b_{\eta,\alpha,\beta,j}(\bar A_j,\bar L_{j-1})\mid\bar A_j,\bar L_{j-1}\},$$

$$S(\alpha)=\sum_{j=1}^J\partial\log f(A_j\mid\bar A_{j-1},\bar L_{j-1};\alpha)/\partial\alpha,\ M(\alpha,\beta)=\omega_J(\alpha)b(Z)\{Y-h(Z;\beta)\},$$

$$\hat I_k=\partial\phi_k(\beta)/\partial\beta^{\rm T}|_{\hat\beta_{\rm or}}E_n\{\partial M(\hat\alpha,\beta)/\partial\beta^{\rm T}|_{\hat\beta_{\rm or}}\}^{-1},$$

$$\hat\rho_\eta^{\rm T}=E_n[\{M(\hat\alpha,\hat\beta_{\rm or})-U(\eta,\hat\alpha,\hat\beta_{\rm or})\}S(\hat\alpha)^{\rm T}][E_n\{S(\hat\alpha)S(\hat\alpha)^{\rm T}\}]^{-1}.$$

Step 3. Compute the maximum likelihood estimator $(\tilde\alpha,\tilde\delta)$ of (α,δ) in the augmented drop-out models $(j=1,\ldots,J)$

$$f(A_j\mid\bar A_{j-1},\bar L_{j-1};\alpha,\delta)=C_j(\alpha,\delta)f(A_j\mid\bar A_{j-1},\bar L_{j-1};\alpha)\exp\left\{\sum_{k=1}^{K+1}\delta_{k,j}^{\rm T}u_{k,j}(\bar A_j,\bar L_{j-1})\right\},$$

where $u_{k,j}(\bar A_j,\bar L_{j-1})=U_j(\tilde\eta_k,\hat\alpha,\hat\beta_{\rm or})$, $u_{K+1,j}(\bar A_j,\bar L_{j-1})=U_j(\hat\eta_{\rm or},\hat\alpha,\hat\beta_{\rm or})$ and $C_j(\alpha,\delta)=c_j(\bar A_{j-1},\bar L_{j-1};\alpha,\delta)$ is the normalizing constant. Estimate η by $\tilde\eta_{\rm or}$ computed just like $\hat\eta_{\rm or}$ but with $\omega_j(\hat\alpha)$ replaced with $\omega_j(\tilde\alpha,\tilde\delta)$. Finally, compute $\hat\beta_{\rm dr}=\hat\beta_{\rm reg}(\tilde\eta_{\rm or})$ solving (4) with $m(L;\eta)$ re-defined as $m_1(\bar L_0;\eta)$ and with $\eta=\tilde\eta_{\rm or}$.

An argument similar to that in § 2·2 shows that under regularity conditions, the estimator $\hat\beta_{\rm dr}$ from the preceding algorithm satisfies Properties 4–8 with models (24) and (22) instead of (5) and (3) and with $m_1(\bar L_0;\eta)$ instead of $m(L;\eta)$ in equation (4). The key points are: $\hat\beta_{\rm dr}$ is consistent for β_0 and asymptotically normal under model (22) because it is of the form $\hat\beta_{\rm reg}(\tilde\eta_{\rm or})$, and $\tilde\eta_{\rm or}$ is consistent for η_0 and asymptotically normal under (22). The estimator $\hat\beta_{\rm dr}$ is also consistent for β_0 and asymptotically normal under model (24) because it is equal to $\hat\beta\{\tilde\eta_{\rm or},(\tilde\alpha,\tilde\delta)\}$ solving equation (25) since, as argued in the Appendix, (25) is the same as

$$0=E_n[b(Z)\{m_1(\bar L_0;\eta)-h(Z;\beta)\}]+E_n\left(b(Z)\left[\sum_{j=1}^{J-1}\omega_j(\alpha)\{m_{j+1}(\bar L_j;\eta)-m_j(\bar L_{j-1};\eta)\}+\omega_J(\alpha)\{Y-m_J(\eta)\}\right]\right),\quad(26)$$

and the right-hand side evaluates to zero when β,η and α are replaced with $\hat\beta_{\rm dr},\tilde\eta_{\rm or}$ and $(\tilde\alpha,\tilde\delta)$. The rest of the argument follows with the re-definition $S^*_{\delta_k}(\alpha_0,\delta_0)=\{S^*_{\delta_{k,1}}(\alpha_0,\delta_0)^{\rm T},\ldots,$

$S^*_{\delta_{K,J}}(\alpha_0, \delta_0)^{\mathrm{T}}\}^{\mathrm{T}}$ where $S^*_\alpha(\alpha, \delta)$ and $S^*_{\delta_{k,j}}(\alpha, \delta)$ are the scores for α and $\delta_{k,j}$ in models

$$f^*(A_j \mid \bar{A}_{j-1}, \bar{L}_{j-1}; \alpha, \delta) = C^*_j(\alpha, \delta) f(A_j \mid \bar{A}_{j-1}, \bar{L}_{j-1}; \alpha) \times \exp\left\{\sum_{k=1}^{K+1} \delta^{\mathrm{T}}_{k,j} U_{0,j}(\eta^*_k) + \delta^{\mathrm{T}}_{K+1,j} U_j(\eta^*_{\mathrm{or}})\right\} \quad (j = 1, \ldots, J),$$

with $C^*_j(\alpha, \delta) = c^*_j(\bar{A}_{j-1}, \bar{L}_{j-1}; \alpha, \delta)$ the normalizing constant, $U_{0,j}(\eta^*_k) = U_j(\eta^*_k, \alpha_0, \beta_0) = S^*_{\delta_{k,j}}(\alpha_0, \delta_0)$ and $U_j(\eta^*_{\mathrm{or}}) = U_j(\eta^*_{\mathrm{or}}, \alpha_0, \beta_0) = S^*_{\delta_{K+1,j}}(\alpha_0, \delta_0)$.

5. SIMULATION STUDIES

We carried out four simulation experiments to assess the performance of $\hat{\beta}_{\mathrm{dr}}$ for estimation of $\beta_0 = E(Y)$ with sample sizes $n = 200, 1000$. In each experiment, we generated 1000 Monte Carlo datasets and computed, in addition to $\hat{\beta}_{\mathrm{dr}}$, the estimators $\hat{\beta}_{\mathrm{reg}}(\hat{\eta})$, $\hat{\beta}(\hat{\eta}, \hat{\alpha})$, $\hat{\beta}_{\mathrm{ipw}}$ and $\hat{\beta}_{\mathrm{Cao}}$, the estimator called $\hat{\mu}_{\mathrm{PROJ}}$ in Cao et al. (2009).

In the first two experiments, we generated data as in Kang & Schafer (2007): $(L_1, \ldots, L_4, \varepsilon) \sim N(0, I_5)$ where I_5 is the 5×5 identity matrix, $Y = 210 + 27{\cdot}4L_1 + 13{\cdot}7\sum_{s=2}^4 L_s + \varepsilon$ and $A \sim \mathrm{Ber}\{\mathrm{expit}(-L_1 + 0{\cdot}5L_2 - 0{\cdot}25L_3 - 0{\cdot}1L_4)\}$. As in Kang & Schafer (2007), we computed $X = (X_1, \ldots, X_4)$, where $X_1 = \exp(L_1/2)$, $X_2 = L_2/\{1 + \exp(L_1)\} + 10$, $X_3 = (L_1L_3/25 + 0{\cdot}6)^3$ and $X_4 = (L_2 + L_4 + 20)^2$. In the first experiment, we assumed that the observed data were (A, AY, L). We computed the estimators using the same outcome and missingness models as in Kang & Schafer (2007). The first, correctly specified, outcome and missingness models were, respectively, an additive linear regression of Y on L with intercept and a logistic regression with intercept, covariate L and outcome A. The second models, incorrectly specified, were as the first ones, but with covariate X instead of L. In the second experiment, as in Robins et al. (2007), we recoded A as $1 - A$ and replicated the first experiment. We conducted this experiment because Robins et al. (2007) noted that the favourable performance of $\hat{\beta}_{\mathrm{reg}}(\hat{\eta})$ compared with $\hat{\beta}(\hat{\eta}, \hat{\alpha})$ reported by Kang & Schafer (2007) was reversed when the observed data were $\{1 - A, (1 - A)Y, L\}$; thus, our study includes scenarios favourable to $\hat{\beta}_{\mathrm{reg}}(\hat{\eta})$ and to $\hat{\beta}(\hat{\eta}, \hat{\alpha})$.

In the last two experiments, we generated L, X and A as in the first experiment, but $Y \sim \mathrm{Ber}\{\mathrm{expit}(-60 + 27{\cdot}4L_1 + 13{\cdot}7\sum_{s=2}^4 L_s)\}$ yielding $E(Y) = 0{\cdot}0496$. We purposely chose to simulate a rare outcome Y as we wanted to examine the performance of $\hat{\beta}_{\mathrm{dr}}$ in a setting where $\hat{\beta}(\hat{\eta}, \hat{\alpha})$ had some nontrivial probability of falling outside the parameter space. Our estimators used the same correct and incorrect missingness models as in the first experiment and two logistic regression models for Y: the first, correctly specified, with intercept and covariate L and the second, incorrectly specified, with intercept and covariate X. The fourth experiment differed from the third in that A was recoded as $1 - A$.

The estimator $\hat{\eta}$ was the ordinary least squares estimator and the standard logistic regression estimator in the first two and last two experiments, respectively. The estimators $\hat{\eta}_{\mathrm{or}}$ and $\tilde{\eta}_{\mathrm{or}}$ were computed as $\hat{\eta}$ except that each unit was weighted by the inverse of specific estimates of the missingness probabilities as described in Steps 1 and 3 of the procedure in § 2·2.

Tables 1 and 2 report results for continuous and binary Y, respectively, and provide Monte Carlo estimates of the bias, root mean square error and median absolute error of the estimators of β_0. Bootstrap estimators of their Monte Carlo standard errors can be found in Tables 3 and 4 of the Supplementary Material.

Table 1. *Monte Carlo study of the performance of the proposed estimator with outcome normally distributed and missing at random*

	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE
	Miss-C, OR-C			Miss-I, OR-C			Miss-C, OR-I			Miss-I OR-I		
	$n = 200$, Y observed iff $A = 1$											
$\hat{\beta}_{\text{reg}}(\hat{\eta})$	0·07	2·54	1·77	0·07	2·54	1·77	−0·51	3·38	2·36	−0·51	3·38	2·36
$\hat{\beta}(\hat{\eta}, \hat{\alpha})$	0·07	2·54	1·78	0·07	2·60	1·79	0·38	3·50	2·30	−7·71	44·28	3·59
$\hat{\beta}_{\text{ipw}}$	−0·09	4·16	2·54	2·07	10·99	3·20	−0·09	4·16	2·54	2·07	10·29	3·20
$\hat{\beta}_{\text{Cao}}$	0·07	2·54	1·79	0·06	2·53	1·78	0·08	2·59	1·82	−0·41	3·51	2·14
$\hat{\beta}_{\text{dr}}$	0·07	2·54	1·78	0·06	2·54	1·77	0·33	2·92	2·05	−1·71	3·58	2·42
	$n = 1000$, Y observed iff $A = 1$											
$\hat{\beta}_{\text{reg}}(\hat{\eta})$	0·01	1·19	0·82	0·01	1·19	0·82	−0·82	1·70	1·19	−0·82	1·70	1·19
$\hat{\beta}(\hat{\eta}, \hat{\alpha})$	0·01	1·19	0·82	0·03	1·20	0·81	0·08	1·59	1·08	−9·78	23·55	5·24
$\hat{\beta}_{\text{ipw}}$	−0·01	1·68	1·14	4·40	9·20	2·55	−0·01	1·68	1·14	4·40	9·20	2·55
$\hat{\beta}_{\text{Cao}}$	0·01	1·19	0·81	0·01	1·19	0·81	0·04	1·19	0·84	−1·26	1·81	1·34
$\hat{\beta}_{\text{dr}}$	0·01	1·19	0·81	0·01	1·19	0·81	0·06	1·22	0·86	−2·56	2·91	2·54
	$n = 200$, Y observed iff $A = 0$											
$\hat{\beta}_{\text{reg}}(\hat{\eta})$	0·07	2·54	1·76	0·07	2·54	1·76	5·01	5·79	5·02	5·01	5·79	5·02
$\hat{\beta}(\hat{\eta}, \hat{\alpha})$	0·07	2·54	1·75	0·07	2·54	1·76	0·53	3·83	2·38	3·25	4·59	3·44
$\hat{\beta}_{\text{ipw}}$	0·29	3·89	2·47	3·85	5·02	4·09	0·29	3·89	2·47	3·85	5·02	4·09
$\hat{\beta}_{\text{Cao}}$	0·07	2·55	1·79	0·08	2·54	1·77	0·47	2·61	1·81	0·94	3·21	2·26
$\hat{\beta}_{\text{dr}}$	0·08	2·54	1·76	0·08	2·54	1·77	1·16	3·01	2·03	2·54	3·89	2·82
	$n = 1000$, Y observed iff $A = 0$											
$\hat{\beta}_{\text{reg}}(\hat{\eta})$	0·01	1·19	0·83	0·01	1·19	0·83	4·93	5·10	4·93	4·93	5·10	4·93
$\hat{\beta}(\hat{\eta}, \hat{\alpha})$	0·01	1·19	0·83	0·01	1·19	0·83	0·18	1·67	1·15	3·09	3·40	3·09
$\hat{\beta}_{\text{ipw}}$	0·12	1·68	1·19	3·71	3·98	3·71	0·12	1·68	1·19	3·71	3·98	3·71
$\hat{\beta}_{\text{Cao}}$	0·01	1·19	0·83	0·01	1·19	0·84	0·14	1·21	0·84	1·12	1·70	1·24
$\hat{\beta}_{\text{dr}}$	0·01	1·19	0·83	0·01	1·19	0·84	0·29	1·29	0·94	1·47	2·07	1·54

RMSE, root mean square error; MAE, median absolute error; Miss-C and Miss-I (OR-C and OR-I), correct and incorrect missingness (outcome regression); $\hat{\beta}_{\text{reg}}(\hat{\eta})$, outcome regression estimator; $\hat{\beta}(\hat{\eta}, \hat{\alpha})$, standard double robust estimator; $\hat{\beta}_{\text{ipw}}$ inverse probability weighted estimator; $\hat{\beta}_{\text{Cao}}$, Cao et al. estimator, $\hat{\beta}_{\text{dr}}$, new double robust estimator.

According to theory, when the outcome model is incorrect and the missingness model is correct, $\hat{\beta}(\hat{\eta}, \hat{\alpha})$, $\hat{\beta}_{\text{ipw}}$, $\hat{\beta}_{\text{dr}}$ and $\hat{\beta}_{\text{Cao}}$ are consistent, $\hat{\beta}_{\text{reg}}(\hat{\eta})$ is inconsistent, and both $\hat{\beta}_{\text{dr}}$ and $\hat{\beta}_{\text{Cao}}$ are asymptotically more efficient than $\hat{\beta}(\hat{\eta}, \hat{\alpha})$ and $\hat{\beta}_{\text{ipw}}$. The performance observed for $n = 1000$, as quantified by bias and mean squared error, agrees with the asymptotic results except that when Y is binary, $\hat{\beta}(\hat{\eta}, \hat{\alpha})$, $\hat{\beta}_{\text{ipw}}$, $\hat{\beta}_{\text{dr}}$ and $\hat{\beta}_{\text{Cao}}$ are slightly biased. The comparison of the relative biases of $\hat{\beta}_{\text{dr}}$ and $\hat{\beta}_{\text{Cao}}$ depends on the data generating mechanism. For $A = 0$ and Y continuous, the bias of $\hat{\beta}_{\text{Cao}}$ is smaller than that of $\hat{\beta}_{\text{dr}}$. On the other hand, for Y binary, $\hat{\beta}_{\text{dr}}$ has substantially smaller bias than $\hat{\beta}_{\text{Cao}}$. Although not predicted by theory, when the missingness model is incorrect and the outcome model is correct, all double-robust estimators perform as well as $\hat{\beta}_{\text{reg}}(\hat{\eta})$ when $n = 1000$. When both models were incorrect $\hat{\beta}_{\text{Cao}}$ outperformed $\hat{\beta}_{\text{dr}}$ in Table 1; however, $\hat{\beta}_{\text{dr}}$ outperformed $\hat{\beta}_{\text{Cao}}$ in Table 2. It is not surprising and, indeed predicted by asymptotic theory, that their relative performance would vary with the data generating mechanism. Results for $n = 200$ were qualitatively similar, except that $\hat{\beta}_{\text{Cao}}$ had smaller root mean squared error for Y continuous when the missingness model was correct and the outcome regression model was incorrect.

Finally, when $n = 200$, $A = 1$ and Y binary, of the 1000 replications, $\hat{\beta}(\hat{\eta}, \hat{\alpha})$ fell below zero a total of 27, 54, 27 and 56 times, and $\hat{\beta}_{\text{Cao}}$ fell below zero 49, 94, 59 and 87 times, respectively,

Table 2. *Monte Carlo study of the performance of the proposed estimator with binary outcome missing at random*

	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE
	Miss-C, OR-C			Miss-I, OR-C			Miss-C, OR-I			Miss-I OR-I		
$n = 200$, Y observed iff $A = 1$												
$\hat{\beta}_{\text{reg}}(\hat{\eta})$	−0·42	2·92	1·96	−0·42	2·92	1·96	0·06	2·90	1·95	0·06	2·90	1·95
$\hat{\beta}(\hat{\eta}, \hat{\alpha})$	−0·42	2·92	1·96	−0·42	2·92	1·96	0·07	2·92	1·96	0·04	2·96	1·95
$\hat{\beta}_{\text{ipw}}$	−0·24	4·01	2·56	2·55	9·67	3·31	−0·24	4·01	2·56	2·55	9·67	3·31
$\hat{\beta}_{\text{Cao}}$	−0·21	3·04	1·99	−0·25	3·04	1·96	0·05	2·92	1·98	0·00	2·93	1·96
$\hat{\beta}_{\text{dr}}$	−0·41	2·93	1·96	−0·39	2·93	1·96	0·05	2·90	1·96	0·02	2·91	1·95
$n = 1000$, Y observed iff $A = 1$												
$\hat{\beta}_{\text{reg}}(\hat{\eta})$	−0·02	0·98	0·64	−0·02	0·98	0·64	0·27	1·04	0·67	0·27	1·04	0·67
$\hat{\beta}(\hat{\eta}, \hat{\alpha})$	−0·01	0·95	0·64	0·00	0·95	0·64	0·17	1·09	0·75	−0·42	1·54	0·86
$\hat{\beta}_{\text{ipw}}$	−0·12	1·74	1·11	5·51	11·80	2·89	−0·12	1·74	1·11	5·51	11·80	2·89
$\hat{\beta}_{\text{Cao}}$	0·07	1·08	0·74	0·08	1·26	0·75	0·25	1·07	0·72	0·19	1·12	0·72
$\hat{\beta}_{\text{dr}}$	−0·02	0·95	0·64	−0·01	0·95	0·64	0·06	1·01	0·71	0·04	1·00	0·67
$n = 200$, Y observed iff $A = 0$												
$\hat{\beta}_{\text{reg}}(\hat{\eta})$	0·05	1·65	1·04	0·05	1·65	1·04	0·87	2·24	1·46	0·87	2·24	1·46
$\hat{\beta}(\hat{\eta}, \hat{\alpha})$	0·05	1·65	1·04	0·02	1·65	1·04	0·78	2·19	1·46	0·83	2·23	1·46
$\hat{\beta}_{\text{ipw}}$	−0·02	1·75	1·25	−0·02	1·80	1·25	−0·02	1·75	1·25	−0·02	1·80	1·25
$\hat{\beta}_{\text{Cao}}$	0·07	1·65	1·04	0·06	1·65	1·04	0·89	2·23	1·46	0·92	2·30	1·46
$\hat{\beta}_{\text{dr}}$	0·05	1·65	1·04	0·05	1·65	1·04	0·82	2·18	1·43	0·84	2·20	1·44
$n = 1000$, Y observed iff $A = 0$												
$\hat{\beta}_{\text{reg}}(\hat{\eta})$	−0·03	0·71	0·47	−0·03	0·71	0·47	0·39	0·88	0·59	0·39	0·88	0·59
$\hat{\beta}(\hat{\eta}, \hat{\alpha})$	−0·03	0·71	0·48	−0·03	0·71	0·48	0·11	0·85	0·56	0·25	0·84	0·56
$\hat{\beta}_{\text{ipw}}$	−0·02	0·80	0·56	−0·02	0·84	0·56	−0·02	0·80	0·56	−0·02	0·84	0·56
$\hat{\beta}_{\text{Cao}}$	−0·02	0·72	0·49	−0·03	0·73	0·51	0·30	0·86	0·57	0·44	0·97	0·61
$\hat{\beta}_{\text{dr}}$	−0·04	0·71	0·48	−0·04	0·71	0·49	0·15	0·78	0·52	0·26	0·84	0·55

RMSE, root mean square error; MAE, median absolute error. All figures in the table are multiplied by 100. Miss-C and Miss-I (OR-C and OR-I), correct and incorrect missingness (outcome regression); $\hat{\beta}_{\text{reg}}(\hat{\eta})$, outcome regression estimator; $\hat{\beta}(\hat{\eta}, \hat{\alpha})$, standard double robust estimator; $\hat{\beta}_{\text{ipw}}$, inverse probability weighted estimator; $\hat{\beta}_{\text{Cao}}$, Cao et al. estimator; $\hat{\beta}_{\text{dr}}$, new double robust estimator.

under the four possible scenarios combining correct and incorrect specifications of the missingness and outcome regression models, the first with both correct, the second with the missingness model incorrect and the outcome model correct, the third with the missingness model correct and the outcome model incorrect and the last with both incorrect models. In all other cases, $\hat{\beta}(\hat{\eta}, \hat{\alpha})$ or $\hat{\beta}_{\text{Cao}}$ fell between 0 and 1.

6. CONCLUDING REMARKS

The proposal in this paper relies on the key observation that under the missing at random or the no-unmeasured confounders assumption, efficient estimation of parameters of increasingly larger models for the missingness or treatment probabilities improves the efficiency with which the parameters of models for the full or counterfactual data are estimated. As such, the present proposal can be extended, along the lines of § 4, to the estimation of the parameters of marginal structural mean models and of structural nested mean models for time-dependent treatments in longitudinal studies with time-dependent confounders (Robins, 1999). This extension will be reported elsewhere.

Acknowledgement

Andrea Rotnitzky, Lei Gomez and James Robins were funded by grants from the National Institutes of Health, U.S.A. Andrea Rotnitzky is also affiliated with the Harvard School of Public Health. Mariela Sued was funded by grants from the Agencia de Promocion Cientifica y Tecnica de Argentina and the Consejo Nacional de Investigaciones Cientificas y Tecnicas de Argentina. The authors wish to thank the reviewers for helpful comments.

Supplementary material

Supplementary material available at *Biometrika* online includes the programme used to run the simulation study and tables with bootstrap estimators of the Monte Carlo standard errors of the reported bias, root mean squared error and mean absolute deviation.

Appendix

Proof of the equivalence of equations (7) *and* (14)*, and of equations* (25) *and* (26). To prove the equivalence of (25) and (26), first note that $G^b_{\eta,\alpha,\beta,j} - E_\alpha(G^b_{\eta,\alpha,\beta,j} \mid \bar{A}_{j-1}, \bar{L}_{j-1}) = b(Z)\{m_j(\bar{L}_{j-1};\eta) - h(Z;\beta)\}\{\omega_j(\alpha) - \omega_{j-1}(\alpha)\}$. Replacing $G^b_{\eta,\alpha,\beta,j} - E_\alpha(G^b_{\eta,\alpha,\beta,j} \mid \bar{A}_{j-1}, \bar{L}_{j-1})$ with this expression in equation (25) and rearranging the terms of the sums in the left-hand side of (25) we arrive at equation (26). The equivalence between (7) and (14) is the special case of the equivalence between (25) and (26) when $J = 1$. □

Sketch of the proof that the estimators $\tilde{\eta}_{\text{or}}$ *and* $\hat{\eta}_{\text{or}}$ *converge in probability to the same limit.* This follows because $\hat{\eta}_{\text{or}}$ and $\tilde{\eta}_{\text{or}}$ solve the same system of q estimating equations except that to compute $\tilde{\eta}_{\text{or}}$, $\omega(\hat{\alpha})$ is replaced by $\omega(\tilde{\alpha}, \tilde{\delta})$ in the last p-equations (10) and second, the left-hand side of (10) evaluated at either $\omega(\hat{\alpha})$ or $\omega(\tilde{\alpha}, \tilde{\delta})$ converges in probability to the same expectation. □

References

Bang, H. & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 692–972.

Cao, W., Tsiatis, A. & Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96**, 723–34.

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

Kang, D. Y. L. & Schafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion and rejoinder). *Statist. Sci.* **22**, 523–80.

Robins, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. *Statistical Models in Epidemiology: The Environment and Clinical Trials*, Ed. M. E. Halloran and D. Berry, pp. 95–134, Institute for Mathematics and its Applications 116. New York: Springer.

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. *Proc. Am. Statist. Assoc. Sect. Bayesian Statist. Sci.*, 6–10.

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. & Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87**, 113–24.

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**, 846–66.

Robins, J. M., Gomez, Q., Sued, M. & Rotnitzky, A. (2007). Performance of double-robust estimators when inverse probability weights are highly variable. *Statist. Sci.* **22**, 544–59.

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

Rubin, D. & van der Laan, M. J. (2008). Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. *Int. J. Biostatist.* **4**, article 5.

Scharfstein, D. O., Rotnitzky, A. & Robins, J. M. (1999). Adjusting for non-ignorable drop-out using semiparametric non-response models. *J. Am. Statist. Assoc.* **94**, 1096–20.

TAN, Z. (2006). A distributional approach for causal inference using propensity scores. *J. Am. Statist. Assoc.* **101**, 1619–37.
TAN, Z. (2007). Understanding OR, PS and DR. *Statist. Sci.* **22**, 560–8.
TAN, Z. (2008). Comment: improved local efficiency and double robustness. *Int. J. Biostatist.* **4**, article 10.
TAN, Z. (2010a). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97**, 661–82.
TAN, Z. (2010b). Nonparametric likelihood and doubly robust estimating equations for marginal and nested structural models. *Can. J. Statist.* **38**, 609–32.
VAN DER LAAN, M. J. (2010). Targeted maximum likelihood based causal inference: part I. *Int. J. Biostatist.* **6**, article 2.
VAN DER LAAN, M. J. & ROBINS, J. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.
VAN DER LAAN, M. J. & RUBIN, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostatist.* **2**, article 11.
VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

[*Received July* 2010. *Revised January* 2011]