

James M. Robins (*Harvard School of Public Health, Boston*) and Naisyin Wang (*Texas A&M University, College Station*)

We describe a reanalysis of the five data sets of Clayton *et al.* using recent methods for obtaining locally efficient estimators in semiparametric models with missing data (Robins and Rotnitzky, 1992; Robins *et al.*, 1994; Robins and Ritov, 1997). Following Clayton *et al.*, our goal is to estimate the logistic incidence model

$$\text{logit}(\text{pr}[D_1 = 1|X, D_0 = 0]) = \phi'X, \quad (1)$$

based on independent realizations $O_i, i = 1, \dots, 10000$, of the observed data $O = (\Delta_0, \Delta_1, \Delta_0 D_0, \Delta_1 D_1, S_0, S_1, X)$ where $\phi'X = \phi'_{\text{age}}X_{\text{age}} + \phi'_{\text{sex}}X_{\text{sex}}$, and, for $j \in \{0, 1\}$, $\Delta_j = 1$ if D_j is observed and $\Delta_j = 0$ otherwise, and S_j is a surrogate for D_j . Furthermore, by design, the selection probabilities

$$\pi(k, l) \equiv \text{pr}(\Delta_0 = k, \Delta_1 = l|X, S_1, S_0, D_1, D_0) = \text{pr}(\Delta_0 = k, \Delta_1 = l|S_1, S_0, X) \quad (2)$$

are known.

In this commentary, we consider the semiparametric model characterized by equations (1) and (2) in which the nuisance densities $f(S_1, S_0|X, D_0, D_1)$, $f(D_0|X)$ and $f(X)$ are considered of no scientific interest and are left completely unrestricted. We refer to estimators of ϕ which are guaranteed to be uniformly asymptotically normal and asymptotically unbiased, whatever the nuisance densities are, as semiparametric estimators. The parametric maximum likelihood (ML) and parametric mean score imputation (MSI) estimators of Clayton *et al.* are not semiparametric and may be inconsistent for ϕ with an unknown degree of bias. The inverse probability weighting (IPW) estimators of Clayton *et al.* are semiparametric but are very inefficient. This raises the question of whether more efficient semiparametric estimators exist. Robins *et al.* (1994) proved that

- (a) any semiparametric estimator of ϕ based on the full data $F_i, i = 1, \dots, 10000$, $F \equiv (W, D_1, D_0)$ with $W = (X, S_0, S_1)$ is asymptotically equivalent to an estimator $\hat{\phi}(h)$ solving

$$0 = \sum_i h(X_i) \epsilon_i(\phi)(1 - D_{0i}),$$

where $\epsilon(\phi) = D_1 - \text{expit}(\phi'X)$, $\text{expit}(u) = \exp(u)/(1 + \exp(u))$ and $h(X)$ is a vector function of the dimension of ϕ , and

- (b) any semiparametric estimator of ϕ based on the observed data $O_i, i = 1, \dots, 10000$, is asymptotically equivalent to an estimator solving an augmented IPW estimating equation $0 = \Sigma_i U_i(h, q, \phi)$ where

$$U(h, q, \phi) = \Delta_1 \Delta_0 h(X) \epsilon(\phi)(1 - D_0)/\pi(1, 1) - A(q),$$

$$A(q) = \Delta_1 \Delta_0 E[q(O)|F]/\pi(1, 1) - q(O),$$

and $q(O)$ is a vector function of $\text{dim}(\phi)$.

For any function

$$q(O) \equiv \Delta_0 \Delta_1 q_{11}(W, D_1, D_0) + \Delta_0(1 - \Delta_1)q_{10}(W, D_0) \\ + (1 - \Delta_0)\Delta_1 q_{01}(W, D_1) + (1 - \Delta_0)(1 - \Delta_1)q_{00}(W),$$

the conditional expectation

$$E[q(O)|F] = \pi(1, 1)q_{11}(W, D_1, D_0) + \pi(1, 0)q_{10}(W, D_0) + \pi(0, 1)q_{01}(W, D_1) + \pi(0, 0)q_{00}(W)$$

can be explicitly computed using the known $\pi(k, l)$. Note that, given $h(X)$ and $q(O)$, $U(h, q)$ subtracts a function $A(q)$ that has conditional mean 0 given the full data F from the IPW estimating function that uses $h(X)$ in place of Clayton *et al.*'s X . The most efficient estimator $\hat{\phi}(h_{\text{eff}}, q_{\text{eff}})$ in the class $\{\phi(h, q)\}$ has asymptotic variance equal to the semiparametric variance bound for the model. The semiparametric variance bound V is the supremum over all correctly specified fully parametric models for the nuisance densities of the parametric Cramér–Rao variance bounds for ϕ .

With non-monotone missing data, the optimal choices $h_{\text{eff}}(X)$ and $q_{\text{eff}}(O)$ are calculated as follows (Robins and Rotnitzky (1992), Robins *et al.* (1994), propositions 8.1 and 8.3, and van der Laan (1993)). Given any vector function $B = b(F)$ of $\dim(\phi)$, define

$$q(O, b) \equiv E[B|O] \equiv \Delta_1 \Delta_0 B + \Delta_0(1 - \Delta_1)E[B|D_0, W] \\ + \Delta_1(1 - \Delta_0)E[B|D_1, W] + (1 - \Delta_0)(1 - \Delta_1)E[B|W].$$

Define $h(X, b) = E[m(B)|X, D_0 = 0, D_1 = 1] - E[m(B)|X, D_0 = 0, D_1 = 0]$ where $m(B) = E[q(O, b)|F]$. Then $q_{\text{eff}}(O) = q(O, b^\infty)$ and $h_{\text{eff}}(X) = h(X, b^\infty)$ where $B^\infty = b^\infty(F)$ is the value at convergence of the following iterative algorithm. Set $B^0 \equiv \mathbf{0}$; then

$$B^{k+1} = B^k + X \epsilon(\phi)(1 - D_0) - m(B^k) + E\{[m(B^k) \\ - B^k] \epsilon(X, D_0 = 0) [\text{expit}(\phi'X)\{1 - \text{expit}(\phi'X)\}]^{-1} \epsilon(\phi)(1 - D_0)\}.$$

To implement this algorithm, we require the joint distribution of the full data F . Hence, in practice, we would

- (a) specify and fit, from the observed data O_i , by ML a fully parametric probability model as did Clayton *et al.* and
- (b) calculate $\hat{h}_{\text{eff}}(X)$ and $\hat{q}_{\text{eff}}(O)$ based on the estimated distribution.

The resulting estimator $\hat{\phi}(\hat{h}_{\text{eff}}, \hat{q}_{\text{eff}})$ is called locally semiparametric efficient at the specified fully parametric model, since

- (a) it will attain the semiparametric variance bound V if the fully parametric model is correct and
- (b) will remain a semiparametric estimator with less but still reasonable efficiency when the parametric models for the nuisance densities are misspecified.

To demonstrate the major efficiency improvements available with nearly correct specification, we calculated h_{eff} and q_{eff} for each of the five data sets by using the empirical joint distribution of the full data $F_i, i = 1, \dots, 10000$ (made available to us by the authors), except with the no age–sex interaction restriction of model (1) imposed. Since X has 12 levels, and, for $j \in \{0, 1\}$, the S_j and D_j have 31 and two levels respectively, the joint distribution of F is discrete with $12 \times 31^2 \times 2^2 = 46128$ potential points of support. Table 1 gives point estimates and standard errors for the semiparametric efficient estimator $\hat{\phi}_{\text{sex}}(h_{\text{eff}}, q_{\text{eff}})$ of ϕ_{sex} . We calculated the standard error of $\hat{\phi}(h_{\text{eff}}, q_{\text{eff}})$ using the usual sandwich variance estimator. Unlike the inefficient IPW estimator, the efficiency of the semiparametric efficient estimator cannot be improved by using empirical estimates of the selection probabilities $\pi(k, l)$.

Table 1. Point estimates and standard errors (in parentheses) for the semiparametric efficient estimator $\hat{\phi}_{\text{sex}}(h_{\text{eff}}, q_{\text{eff}})$

Estimates for the following data sets:					Average standard error
1	2	3	4	5	
–0.100 (0.274)	–0.249 (0.197)	–0.325 (0.235)	–0.544 (0.314)	0.117 (0.209)	0.245

Comparing our results with those of Clayton *et al.* in their Fig. 6 and Table 2, we see (by squaring ratios of standard errors) that by using $\hat{\phi}_{\text{sex}}(h_{\text{eff}}, q_{\text{eff}})$ rather than their IPW estimators we would require somewhere between a fifth and an eighth the sample size to achieve the same confidence interval length without having to impose any additional assumptions beyond equations (1) and (2). Indeed, $\hat{\phi}(h_{\text{eff}}, q_{\text{eff}})$ is nearly as efficient as Clayton *et al.*'s MSI2 and ML estimators.

Our methodology for constructing locally semiparametric efficient augmented IPW estimators in semiparametric models with missingness by design is completely general and fully consistent with the spirit of the design or randomization-based tradition in survey sampling. In contrast with parametric MSI and ML estimators, augmented IPW estimators perform well in moderately sized samples even if X and S are high dimensional and continuous, because augmented IPW estimators protect against bias by using the fact that the sampling probabilities $\pi(k, l)$ are known. Of course, occasionally, even a locally semiparametric efficient estimator may have variability that is too great to enable us to reach important substantive conclusions. In that case, we should report both the locally semiparametric efficient estimator and the more precise ML or parametric MSI estimator (with standard errors) to make the reader aware that the apparent precision of these last two estimators depends on possibly incorrect assumptions about the nuisance densities. Finally, augmented IPW methods can be extended to studies with missingness by happenstance (as well as design) by specifying additional ignorable or non-ignorable parametric models for the probability of missingness by happenstance (Rotnitzky and Robins, 1997).