Published in final edited form as:

Int Stat Rev. 2011 August 1; 79(2): 221–223. doi:10.1111/j.1751-5823.2011.00144.x.

Discussion of “Connections Between Survey Calibration Estimators and Semiparametric Models for Incomplete Data” by T. Lumley, P.A. Shaw & J.Y. Dai

Anastasios A. Tsiatis and Marie Davidian

Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA

We congratulate the authors (henceforth LSD) on a long overdue, detailed review of the connection between the AIPW estimators derived in the incomplete data context via semiparametric theory by Robins, Rotnitzky, and colleagues and the survey calibration estimators used widely in survey sampling. Although this connection has been noted previously (e.g. Robins & Rotnitzky, 1998; Rotnitzky, 2009), the present article appears to be the first in the statistical literature to offer a more comprehensive account. We had only a passing familiarity with this connection, and we are grateful to the editors for the opportunity to offer this discussion, whose preparation required us to acquire a deeper understanding. In what follows, we hope to complement the presentation of LSD by highlighting some further relationships and differences between the two perspectives.

We adopt the notation used by the authors and consider estimation of the population total T . We focus on the regression estimator $\widehat{T}_{reg}$, which, as in Section 2.1 of LSD may be written equivalently as $\widehat{T}(\beta)$, where

$$\widehat{T}(\beta) = \widehat{T} - \sum_{i=1}^N \left(\frac{R_i - \pi_i}{\pi_i} \right) x_i \beta, \quad (1)$$

a representation that may be more familiar to statisticians well-versed in the AIPW literature. We reiterate and expand upon some important differences between the missing data and survey sampling contexts noted by LSD.

In survey sampling, the realizations $(x_1, y_1), \dots, (x_N, y_N)$ that comprise the population are regarded as fixed, and inference on $T = \sum_{i=1}^N y_i$, or, equivalently, the population mean $N^{-1}T$, is the goal. This is based on data $(x_i, R_i, R_i y_i)$, $i = 1, \dots, N$, drawn from the population according to a fixed, known design, where $n = \sum_{i=1}^N R_i$, and $\Pr(R_i = 1) = \pi_i$ and $\Pr(R_i = 1, R_j = 1) = \pi_{ij}$ for π_i known and π_{ij} known or unknown, $i, j = 1, \dots, N$. The (x_i, y_i) may be viewed as realizations of random variables (X_i, Y_i) , $i = 1, \dots, N$, representing an independent and identically distributed (iid) sample from some superpopulation; however interest focuses on the fixed quantity T (Särndal et al., 2003).

In contrast, in the incomplete data context, interest is in estimation of $\mu = E(Y)$, a parameter associated with the superpopulation. Instead of observing a realization of iid (X_i, Y_i) , $i = 1, \dots, N$, we observe a realization of iid $(X_i, R_i, R_i Y_i)$, $i = 1, \dots, N$, where R_i is an indicator of whether or not the value of Y_i is observed or missing. Ordinarily, the probabilities of observing Y_i for each i are not fixed by design; rather, missingness arises according to some unknown mechanism about which some assumption is made. A common assumption is that R_i is conditionally independent of Y_i given X_i , the so-called “missing at

random” (MAR) assumption, under which $\pi(X_i) = \Pr(R_i = 1 | X_i, Y_i) = \Pr(R_i = 1 | X_i)$. MAR cannot be verified from the observed data, so the validity of inference on μ depends on its unknown relevance. Even if MAR is plausible, which we assume henceforth, the function $\pi(X)$ is not known and thus must be estimated based on the observed data, usually via maximum likelihood for a posited parametric model, yielding predicted values $\hat{\pi}_i$.

From the survey sampling perspective, because the π_i are known, $\hat{T}(\beta)$ in (1) is a consistent estimator for T for any fixed β , and the choice $\hat{\beta}$ given in Section 2 of LSD leading to $\hat{T}_{reg}$ is meant to yield an estimator that is more precise than $\hat{T}$. In the incomplete data setting, $N^{-1}\hat{T}(\beta)$ with the $\hat{\pi}_i$ substituted for the π_i need not be consistent for μ unless the model for $\pi(X_i)$ is correctly specified. If this model is correct, then it follows from semiparametric theory that all regular, asymptotically linear estimators for μ may be written in the form (Robins et al., 1994)

$$N^{-1} \sum_{i=1}^N \left\{ \frac{R_i Y_i}{\pi_i} - \left(\frac{R_i - \pi_i}{\pi_i} \right) \phi(X_i) \right\}, \quad (2)$$

where $\phi(X)$ is an arbitrary function of X . The choice of $\phi(X)$ leading to the most precise estimator within class (2) is $\phi(X) = E(Y|X)$. Accordingly, a (usually parametric) model for $E(Y|X)$ may be posited and fitted, and the predicted values substituted for $\phi(X_i)$ in (2).

Scharfstein et al. (1999) made the critical observation that such an estimator is “doubly robust” (DR), i.e. is consistent for μ as long as at least one of the posited models for $\pi(X)$ or $E(Y|X)$ is correct. Because of the protection afforded by this property, DR estimators have been advocated for routine use. In practice, posited models for $E(Y|X)$ might be linear, generalized linear, or arbitrarily nonlinear in a parameter β , depending on the nature of Y . Usually, the posited model is fitted using ordinary or iteratively reweighted least squares based on the pairs (x_i, y_i) for which $R_i = 1$, which, under MAR, would yield a consistent estimator for in a correctly specified model $m(X, \beta)$, say, for $E(Y|X)$.

Kang and Schafer (2007) evaluated the performance of the usual DR estimator for μ in a missing data context under specific simulation scenarios with continuous Y , linear $m(x, \beta) = x\beta$, and β estimated by ordinary least squares. The estimator exhibited poor performance under scenarios where the models for $\pi(X)$ and $E(Y|X)$ were only slightly misspecified and/or when the relative magnitudes of the $\hat{\pi}_i$ were extremely disparate, with $\hat{\pi}_i$ relatively very small for some i , leading Kang and Schafer to issue a strong warning against its routine use. Because of the observational nature of the data, where missingness is by happenstance rather than design, such $\hat{\pi}_i$ may be encountered in practice.

These results led us (Cao et al., 2009; see also Tan, 2006, 2007 and Tsiatis & Davidian, 2007) to speculate that this poor performance may be partly a consequence of the method used to estimate β in the posited model $m(x, \beta)$. We proposed considering the class of DR estimators in (2), where $\phi(X_i)$ is replaced by $m(X_i, \beta)$, and, among such estimators indexed by β , found the value of β that minimizes the variance of estimators within the class when $\pi(X)$ is correctly specified regardless of whether or not $m(X_i, \beta)$ is correct, and a means of estimating this optimal β . The estimator for the optimal β is not ordinary or usual weighted least squares but, rather, involves, in the ideal case where the π_i were known, a weighted regression with weights $(1 - \pi_i)/\pi_i^2$, with modification when π_i are estimated by $\hat{\pi}_i$ as above. Cao et al. (2009) reported simulations showing that the DR estimator incorporating this estimator for the optimal β demonstrated vastly improved performance in the Kang and Schafer and other scenarios. Tsiatis et al. (2011) extended this idea to DR estimators in the more complex setting of longitudinal studies with monotone dropout.

We were interested to learn that the same tactic of finding the optimal β minimizing the variance of estimators for T of the form $\widehat{T}(\beta)$ in (1) and an estimator for this optimal β using a weighting scheme similar in spirit to that in Cao et al. (2009) was proposed by Montanari (1987); see also Berger et al. (2003). Given that in the survey sampling context, with the π_i determined by design, the issue of disparate π_i would not be as pronounced, we suspect that the gains in performance realized by such an approach may not be as dramatic.

We again compliment the authors on an insightful and useful article.

Acknowledgments

This work was supported by grants R37 AI031789, R01 CA051962, R01 CA085848, and P01 CA142538 from the National Institutes of Health.

Additional References

- Berger YG, Tirari MEH, Tillé Y. Towards optimal regression estimation regression in sample surveys. *Aust. & N. Zeal. J. Statist.* 2003; 45:319–329.
- Cao W, Tsiatis AA, Davidian M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika.* 2009; 96:723–734. [PubMed: 20161511]
- Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* 2007; 22:523–539.
- Montanari GE. Post-sampling efficient QR-prediction in large-sample surveys. *Int. Statist. Rev.* 1987; 55:191–202.
- Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* 1994; 89:846–866.
- Rotnitzky, A. Inverse probability weighted methods.. In: Fitzmaurice, G.; Davidian, M.; Vereke, G.; Molenberghs, G., editors. *Longitudinal Data Analysis*. Chapman & Hall/CRC; Boca Raton: 2009. p. 453–476.
- Scharfstein DO, Rotnitzky A, Robins JM. Rejoinder to “Adjusting for nonignorable drop-out using semiparametric nonresponse models.”. *J. Amer. Statist. Assoc.* 1999; 94:1135–1146.
- Tan Z. A distributional approach for causal inference using propensity scores. *J. Amer. Statist. Assoc.* 2006; 101:1619–37.
- Tan Z. Understanding OR, PS and DR. *Statist. Sci.* 2007; 22:560–8.
- Tsiatis AA, Davidian M. Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* 2007; 22:569–573.
- Tsiatis AA, Davidian M, Cao W. Improved doubly robust estimation when data are monotonely coarsened, with application to longitudinal studies with dropout. *Biometrics.* 2011 in press.