Statistical Science
2007, Vol. 22, No. 4, 523–539
DOI: 10.1214/07-STS227
© Institute of Mathematical Statistics, 2007

Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data[1]

Joseph D. Y. Kang and Joseph L. Schafer

Abstract. When outcomes are missing for reasons beyond an investigator's control, there are two different ways to adjust a parameter estimate for covariates that may be related both to the outcome and to missingness. One approach is to model the relationships between the covariates and the outcome and use those relationships to predict the missing values. Another is to model the probabilities of missingness given the covariates and incorporate them into a weighted or stratified estimate. Doubly robust (DR) procedures apply both types of model simultaneously and produce a consistent estimate of the parameter if either of the two models has been correctly specified. In this article, we show that DR estimates can be constructed in many ways. We compare the performance of various DR and non-DR estimates of a population mean in a simulated example where both models are incorrect but neither is grossly misspecified. Methods that use inverse-probabilities as weights, whether they are DR or not, are sensitive to misspecification of the propensity model when some estimated propensities are small. Many DR methods perform better than simple inverse-probability weighting. None of the DR methods we tried, however, improved upon the performance of simple regression-based prediction of the missing values. This study does not represent every missing-data problem that will arise in practice. But it does demonstrate that, in at least some settings, two wrong models are not better than one.

Key words and phrases: Causal inference, missing data, propensity score, model-assisted survey estimation, weighted estimating equations.

1. INTRODUCTION

1.1 Purpose

A new class of methods called doubly robust (DR) procedures was designed to mitigate selection bias arising from uncontrolled nonresponse and attrition, nonrandom treatment assignment in observational studies and noncompliance in randomized experiments (Robins and Rotnitzky, 2001; van der Laan and Robins, 2003). DR methods require specification of two models: one that describes the population of responses, and another that describes the process by which the data are filtered or selected to produce the observed sample. The distinguishing feature of DR estimates is that

Joseph D. Y. Kang is Research Associate, The Methodology Center, 204 E. Calder Way, Suite 400, State College, Pennsylvania 16801, USA (e-mail: josephkang@stat.psu.edu). Joseph L. Schafer is Associate Professor, The Methodology Center, 204 E. Calder Way, Suite 400, State College, Pennsylvania 16801, USA (e-mail: jls@stat.psu.edu).

[1]Discussed in 10.1214/07-STS227A, 10.1214/07-STS227B, 10.1214/07-STS227C and 10.1214/07-STS227D; rejoinder at 10.1214/07-STS227REJ.

they remain asymptotically unbiased if one of the two models is misspecified—that is, they consistently estimate their targets if either model is true (Robins and Rotnitzky, 1995).

DR methods are a refinement of a weighted estimating-equations approach to regression with incomplete data proposed by Robins, Rotnitzky and Zhao (1994, 1995) and Rotnitzky, Robins and Scharfstein (1998). Further explanation and evaluation of DR estimators has been given by Lunceford and Davidian (2004), Carpenter, Kenward and Vansteelandt (2006), Davidian, Tsiatis and Leon (2005) and Bang and Robins (2005). What is not widely known, however, is that the methods developed by Robins et al. are not the only way to achieve double robustness. Many other types of estimates possess the DR property. Generalized regression estimators developed for sample surveys (Cassel, Särndal and Wretman, 1977), also known as model-assisted survey estimators (Särndal, Swensson and Wretman, 1989, 1992), have this property, as does a new class of parametric methods developed by Little and An (2004) and several other methods that do not seem to have been described before.

The first purpose of this article is pedagogical. We review a variety of incomplete-data estimation strategies, describe various ways in which the DR property can arise, and connect the recent articles on DR estimation to similar techniques found in the literature on sample surveys and causal inference. Our second purpose is to investigate the practical behavior of these estimators not only when either of the underlying models is correct, but in a scenario where both models are moderately misspecified.

1.2 Description of the Problem

For simplicity, we focus on the problem of estimating a population mean from an incomplete dataset. Many of the methods we present have been applied to the more general problem of estimating population-average regression coefficients, and we will describe those extensions where appropriate. Let us suppose that we have a random sample of units $i = 1, \ldots, n$ from an infinite population. The variable of primary interest is y_i. Let t_i be the response indicator for y_i, so that $t_i = 1$ if y_i is observed and $t_i = 0$ if y_i is missing. For each unit, there is an observed p-dimensional vector of covariates x_i that may be related both to y_i and to t_i. A schematic representation of the sample data is shown in Figure 1(a). (In this figure, the indices $i = 1, \ldots, n$ have been permuted so that the sample units having $t_i = 1$ appear first.) Denote the numbers

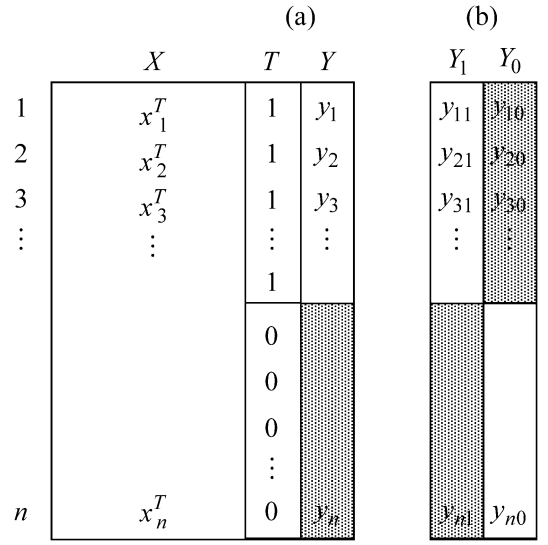


FIG. 1. *Schematic representation of sample data for estimating* (a) *a population mean and* (b) *an average causal effect, with missing values denoted by shading.*

of respondents and nonrespondents by $n^{(1)} = \sum_i t_i$ and $n^{(0)} = \sum_i (1 - t_i)$, the population response and nonresponse rates by $r^{(1)} = P(t_i = 1)$ and $r^{(0)} = P(t_i = 0)$, and the sample rates by $\hat{r}^{(1)} = n^{(1)}/n$ and $\hat{r}^{(0)} = n^{(0)}/n$.

The sample mean of the observed y_i's,

$$\bar{y}^{(1)} = \frac{1}{n^{(1)}} \sum_i t_i y_i,$$

consistently estimates the mean for the respondents, $\mu^{(1)} = E(y_i \mid t_i = 1)$, under any reasonable population distribution and response mechanism. An estimator having this property will be said to be strongly robust. In general, there is no strongly robust estimate of the mean for the nonrespondents, $\mu^{(0)} = E(y_i \mid t_i = 0)$, or for the mean of the entire population, $\mu = E(y_i) = r^{(0)}\mu^{(0)} + r^{(1)}\mu^{(1)}$, based on the observed data alone. Naive estimates such as $\bar{y}^{(1)}$ may work well enough if $\hat{r}^{(0)}$ is small and the relationships among x_i, t_i and y_i are weak. As the nonresponse rate and the strength of these relationships grow, adjusting for selection bias becomes important, and inferences become sensitive to the assumptions underlying the adjustment.

This problem is closely related to estimating an average causal effect from an experiment or observational study. Suppose now that t_i is an indicator of the treatment received by unit i. Associated with unit i is a pair of potential outcomes: the response y_{i1} that is realized if $t_i = 1$, and another response y_{i0} that is realized if $t_i = 0$. This situation is depicted in Figure 1(b).

The causal effect of the treatment on unit i, defined as $y_{i1} - y_{i0}$, is unobservable because one of the two potential outcomes is necessarily missing. It is often of interest to estimate the average causal effect (ACE) in the population,

$$ACE = E(y_{i1}) - E(y_{i0}),$$

or the ACE's among the treated and the untreated,

$$ACE^{(1)} = E(y_{i1} \mid t_i = 1) - E(y_{i0} \mid t_i = 1),$$
$$ACE^{(0)} = E(y_{i1} \mid t_i = 0) - E(y_{i0} \mid t_i = 0).$$

The notion of potential outcomes was introduced by Neyman (1923) for randomized experiments and by Rubin (1974a) for nonrandomized studies. Reviews of causal inference from this perspective are given by Holland (1986), Winship and Sobel (2004), Gelman and Meng (2004) and Rubin (2005). From Figure 1(b), it is apparent that the correlation between y_{i1} and y_{i0} [more precisely, the partial correlation between them given x_i; see Rubin (1974b)] cannot be estimated from the observed data. Without prior information on what this correlation might be—and it is unclear from where such information would come—one may separate the problem of estimating an ACE into independent estimation of the means of y_{i1} and y_{i0}. Any of the methods described in this article can be used to estimate an average causal effect by applying the method separately to each potential outcome. When estimating ACE's, rates of missing information tend to be high and sensitivity to modeling assumptions may be acute.

1.3 Assumptions

Throughout this article, we will assume that the response mechanism is unconfounded in the sense that y_i and t_i are conditionally independent given x_i (Rubin, 1978); this assumption has also been called strong ignorability (Rosenbaum and Rubin, 1983). Under strong ignorability, the joint distribution of the complete data can be written as

$$P(X, T, Y) = \prod_i P(x_i) P(t_i \mid x_i) P(y_i \mid x_i).$$

Strong ignorability implies that the missing y_i's are missing at random (MAR) (Rubin, 1976; Little and Rubin, 2002). In many applications, MAR is unrealistic. Nevertheless, this assumption provides an important benchmark and point of departure for sensitivity analyses, and it is the foundation upon which DR procedures rest.

None of the methods we examine will require an explicit model for the covariates, but they will all make assumptions about $P(t_i \mid x_i)$, $P(y_i \mid x_i)$ or both. Denote the response probability for unit i by

$$P(t_i = 1 \mid x_i) = \pi_i(x_i) = \pi_i. \tag{1}$$

This probability is called the propensity score (Rosenbaum and Rubin, 1983). A proposed functional form for (1) will be called a π-model. If estimates of the propensity scores are needed, they are often taken to be

$$\hat{\pi}_i = \operatorname{expit}(x_i^T \hat{\alpha}) = \frac{\exp(x_i^T \hat{\alpha})}{1 + \exp(x_i^T \hat{\alpha})},$$

where $\hat{\alpha}$ is the maximum-likelihood (ML) estimate of the coefficients from the logistic regression of $t_1, \ldots, t_n$ on $x_1, \ldots, x_n$. In many situations, of course, the assumed form of the π-model is not correct, and this misspecification can be problematic depending on how the $\hat{\pi}_i$'s are used.

Let us also define $E(y_i \mid x_i) = m(x_i) = m_i$, so that

$$y_i = m(x_i) + \varepsilon_i \tag{2}$$

with $E(\varepsilon_i) = 0$. A functional form for $m(x_i)$ will be called a y-model. When an estimate of m_i is needed, an obvious candidate is $\hat{m}_i = x_i^T \hat{\beta}$, where $\hat{\beta}$ is the vector of coefficients from the linear regression of y_i on x_i estimated from the respondents; y-models with nonlinear link functions are also straightforward (McCullagh and Nelder, 1989). In most cases, an analyst's regression model will be only a rough approximation to the true y-model. The implications of this misspecification can be serious if $P(x_i \mid t_i = 1)$ and $P(x_i \mid t_i = 0)$ are very different, because the $\hat{m}_i$'s for the nonrespondents will then be based on extrapolation. This is particularly true when x_i is high-dimensional because of the so-called curse of dimensionality. With many covariates, it becomes difficult to specify a y-model that is sufficiently flexible to capture important nonlinear effects and interactions, yet parsimonious enough to keep the variance of prediction manageably low.

Under ignorability, the π_i's play no role in likelihood-based or Bayesian analyses for the parameters of the y-model (Rubin, 1976). The parametric approach—specifying a full model for y_i and ignoring the π_i's—is emphasized in texts on missing data by Rubin (1987), Schafer (1997), Little and Rubin (2002) and others. Nevertheless, many advocates of the parametric approach also recognize that the π_i's are useful for model validation and criticism (Gelman, Carlin, Rubin and Stern, 2004, Chapters 6–7). On the other hand, much of the literature on causal inference in the tradition of Rosenbaum and Rubin (1983) eschews

models for y_i in favor of matching and stratification based on propensity scores (e.g., Rosenbaum, 2002). The latter is motivated in part by a perceived inability to model the responses well enough to mitigate the dangers of extrapolation. Other uses for propensity scores, including inverse-propensity weighting (Robins, Rotnitzky and Zhao, 1994), also rely heavily on a π-model while relaxing assumptions about y_i. DR methods will require both a y-model and a π-model but remain consistent if one or the other is wrong. As we shall see, however, this property does not necessarily translate into improved performance when both models fail.

1.4 A Simulated Example

Consider the following example which, although artificial, bears some resemblance to what we have encountered in a real study. For each unit $i = 1, \ldots, n$, suppose that $(z_{i1}, z_{i2}, z_{i3}, z_{i4})^T$ is independently distributed as $N(0, I)$ where I is the 4×4 identity matrix. The y_i's are generated as

$$y_i = 210 + 27.4z_{i1} + 13.7z_{i2} + 13.7z_{i3} + 13.7z_{i4} + \varepsilon_i,$$

where $\varepsilon_i \sim N(0, 1)$, and the true propensity scores are

$$\pi_i = \text{expit}(-z_{i1} + 0.5z_{i2} - 0.25z_{i3} - 0.1z_{i4}).$$

This mechanism produces an average response rate of $r^{(1)} = 0.5$, and the means are $\mu = 210.0$, $\mu^{(1)} = 200.0$ and $\mu^{(0)} = 220.0$. The selection bias in this example is not severe; the difference between the mean of the respondents and the mean of the full population is only one-quarter of a population standard deviation. Nevertheless, this difference is large enough to wreak havoc on the performance of the naive estimate $\bar{y}^{(1)}$ when used as the basis for confidence intervals and tests.

In this example, a logistic regression of t_i on the z_{ij}'s would be a correct π-model, and a linear regression of y_i on the z_{ij}'s would be a correct y-model. We will suppose, however, that instead of being given the z_{ij}'s, the covariates actually seen by the data analyst are

$$\begin{aligned} x_{i1} &= \exp(z_{i1}/2), \\ x_{i2} &= z_{i2}/\big(1 + \exp(z_{i1})\big) + 10, \\ x_{i3} &= (z_{i1}z_{i3}/25 + 0.6)^3, \\ x_{i4} &= (z_2 + z_4 + 20)^2. \end{aligned}$$

This implies that $\text{logit}(\pi_i)$ and m_i are linear functions of $\log(x_{i1})$, x_2, $x_1^2 x_2$, $1/\log(x_1)$, $x_3/\log(x_1)$ and $x_4^{1/2}$. Except by divine revelation, it is unlikely that an analyst who sees only x_i would ever formulate a correct π- or y-model. Rather, he or she would naturally be drawn to models that are linear and logistic in the x_{ij}'s, and those incorrect models look trustworthy. To illustrate, we drew a random sample of $n = 200$ units from this population, which happened to produce exactly $n^{(1)} = 100$ respondents and $n^{(0)} = 100$ nonrespondents. Scatterplots of y_i versus x_{ij}, $j = 1, \ldots, 4$, for the 100 respondents are shown in Figure 2. Regressing y_i on the x_{ij}'s yields coefficients for x_{i1}, x_{i3} and x_{i4} that are highly significant and a coefficient for x_{i2} that is nearly significant at the 0.05-level; the prediction is strong ($R^2 = 0.81$), and a plot of residuals versus fitted values reveals no obvious outliers and little evidence of heteroscedasticity or nonlinearity. (In other samples, some higher-order terms such as x_1^2 and x_1x_2, are significant and might be considered for inclusion in the model. Those terms, however, do little to improve the performance of any of the regression-based methods discussed below, and sometimes they are harmful.)

The covariates seen by the analyst are also related to the t_i's. Side-by-side boxplots of the x_{ij}'s for the $t_i = 0$ and $t_i = 1$ groups are shown in Figure 3(a)–(d). Fitting a logistic model to this sample, we find that the coefficients for x_{i1} and x_{i2} are statistically significant, and all of the deviance residuals lie between -1.85 and $+2.51$. Figure 3(e) shows side-by-side boxplots of the linear predictors $\hat{\eta}_i = \text{logit}(\hat{\pi}_i)$. As one would expect, the distributions of $\hat{\pi}_i$ in the two groups are different but not drastically so. To check the appropriateness of the link function, Hinkley (1985) suggested adding $\hat{\eta}_i^2$ as another covariate to see whether it is related to the response; the coefficient for this extra term was not significantly different from zero ($p = 0.20$), so in this sample an analyst would have little reason to alter the link.

Comparing the fitted values of y_i under the true and misspecified y-models, we find that the correlation between them is approximately 0.9. Similarly, the correlation between the $\hat{\eta}_i$'s under the true and misspecified π-models is also about 0.9. This example appears to be precisely the type of situation for which the DR estimators of Robins et al. were developed. By relying on two reasonably good models, one hopes that at least one is close enough to the truth to yield satisfactory results. Indeed, Bang and Robins (2005, Section 2.1) state:

> In our opinion, a DR estimator has the following advantage that argues for its routine use: if either the [y-model] or the [π-model] is nearly correct, then the bias of a DR estimator of μ will be small. Thus, the DR es-

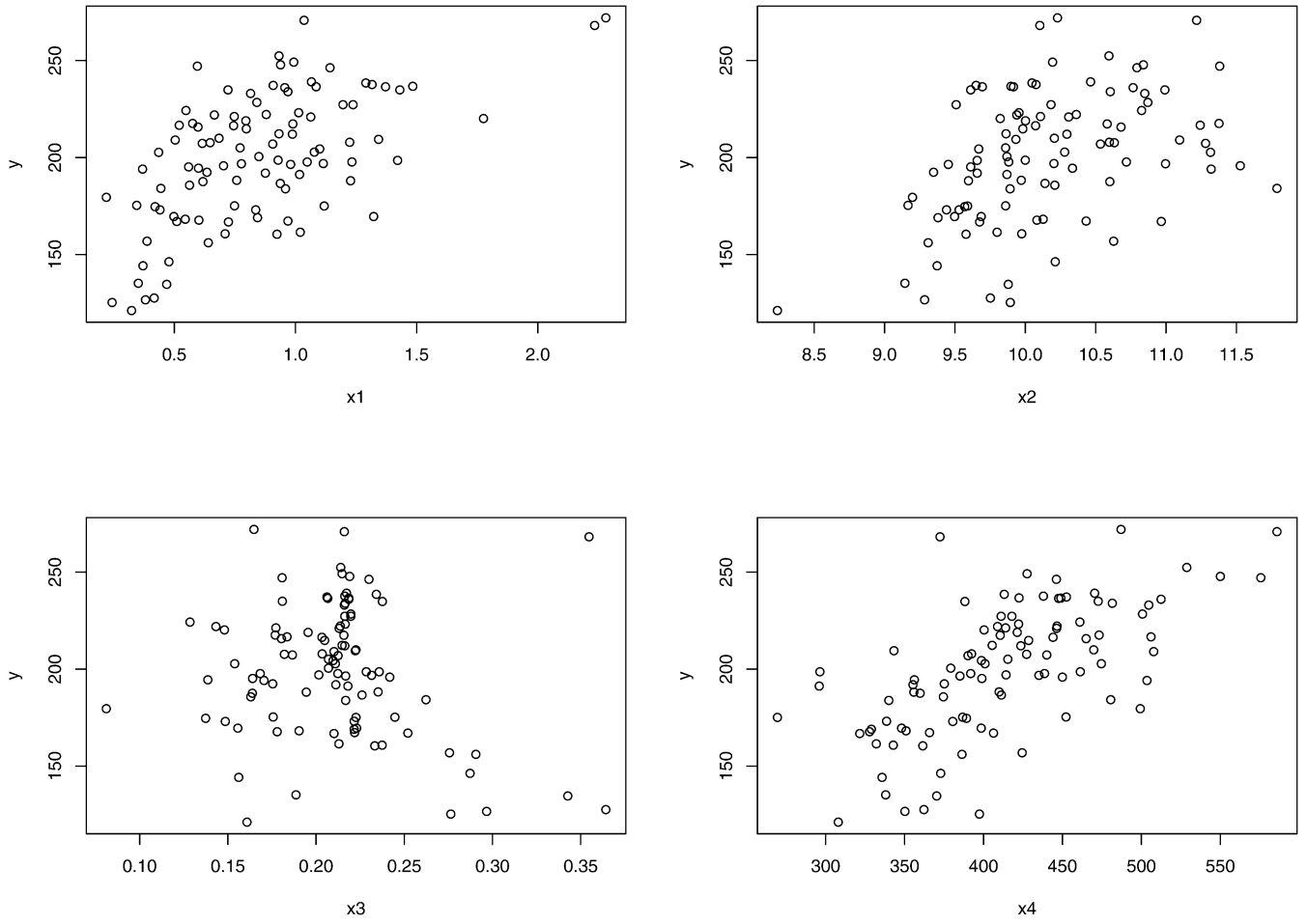


FIG. 2. *Scatterplots of response versus observed covariates for respondents in a sample of* 200 *units.*

timator ... gives the analyst two chances to get nearly correct inferences about the mean of Y.

In Sections 2 and 3, we describe various techniques for estimating μ based on the observed data and evaluate their performance in this simulated example. Some

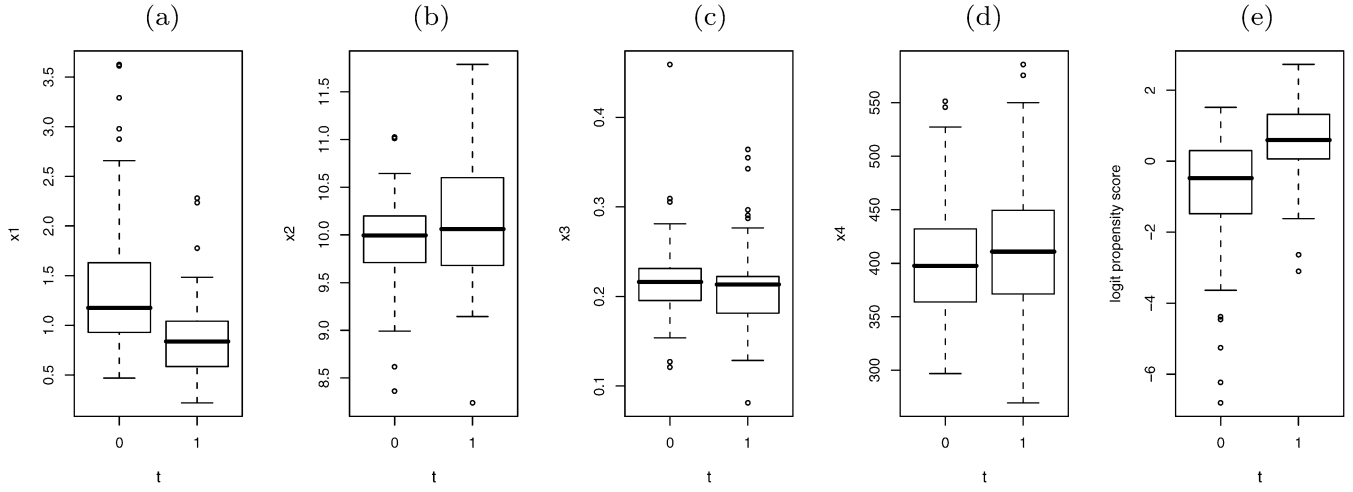


FIG. 3. *Distributions of* (a)–(d) *observed covariates and* (e) *estimated propensity scores for nonrespondents and respondents in a sample of* 200 *units.*

of these methods use a π-model, some use a y-model, and some rely on both. All of the dual-modeling strategies possess a DR property, but they do not perform equally well. Pooling information from two models can be helpful, but the manner in which the information is pooled makes a difference. (Due to space limitations, we will not discuss computation of standard errors. Tractable variance estimates are available for most of these methods, but our purpose is to compare the performance of the estimates themselves.) In Section 4, we will provide further justification for why we constructed our example as we did, and we will outline the crucial differences between this and simulated examples used by Bang and Robins (2005) and others, so that apparently contradictory conclusions can be reconciled.

2. WEIGHTING, STRATIFICATION AND REGRESSION

2.1 Inverse-Propensity Weighting

Weighting observed values by inverse-probabilities of selection was proposed by Horvitz and Thompson (1952) in the context of survey inference for finite populations. The same idea is used in importance sampling, a Monte Carlo technique for approximating the moments of a distribution using random draws from another distribution that approximates it (Hammersley and Handscomb, 1964; Geweke, 1989). It is easy to see that $n^{-1}\sum_i t_i\pi_i^{-1}y_i$, which can be computed from the respondents alone, unbiasedly estimates the mean of the entire population, because strong ignorability implies that

$$\begin{aligned} E(t_i\pi_i^{-1}y_i) &= E(E(t_i\pi_i^{-1}y_i \mid x_i)) \\ &= E(\pi_i\pi_i^{-1}m_i) = \mu. \end{aligned}$$

Precision is often enhanced if we use a denominator of $\sum_i t_i\pi_i^{-1}$ rather than n, so that the estimate becomes a weighted average of the y_i's for the respondents. Normalizing the weights in this manner, and replacing the unknown propensities by estimates derived from a π-model, the inverse-propensity weighted (IPW) estimate becomes

$$\hat{\mu}_{IPW\text{-}POP} = \frac{\sum_i t_i\hat{\pi}_i^{-1}y_i}{\sum_i t_i\hat{\pi}_i^{-1}}. \tag{3}$$

The "*POP*" in the subscript indicates that we are reweighting the respondents to resemble the full population. Alternatively, we may reweight them to approximate the population of nonrespondents (Hirano and Imbens, 2001). An estimate of $\mu^{(0)}$ based on that idea is

$$\hat{\mu}^{(0)}_{IPW\text{-}NR} = \frac{\sum_i t_i\hat{\pi}_i^{-1}(1-\hat{\pi})y_i}{\sum_i t_i\hat{\pi}_i^{-1}(1-\hat{\pi})},$$

and a corresponding estimate of μ is

$$\hat{\mu}_{IPW\text{-}NR} = \hat{r}^{(1)}\bar{y}^{(1)} + \hat{r}^{(0)}\hat{\mu}^{(0)}_{IPW\text{-}NR}. \tag{4}$$

The IPW estimator can also be regarded as the solution to a simple weighted estimating equation $\sum_i w_i U_i = 0$, where $w_i = t_i\hat{\pi}_i^{-1}$ and $U_i = (y_i - \mu)/\sigma^2$ for any $\sigma^2 > 0$. From this standpoint, the IPW method can be generalized to estimate a vector of population-average coefficients for the regression of y_i on an arbitrary set of covariates. In the regression setting, U_i becomes a vector representing the ith unit's contribution to a quasi-score function. Coefficients estimated in this manner are consistent and asymptotically normally distributed provided that the π_i's are bounded away from zero and the π-model has been correctly specified. Asymptotic properties and methods for variance estimation are described by Binder (1983) when the propensities are known, and by Robins, Rotnitzky and Zhao (1994, 1995) when the propensities have been estimated.

In the original method of Horvitz and Thompson (1952), the π_i's were determined by a known survey design. Surveys are usually designed to ensure that IPW estimates are acceptably precise, but the forces of nature that govern uncontrolled nonresponse are often unkind. In missing-data problems, IPW methods assign large weights to respondents who closely resemble nonrespondents, causing the estimates to have high variance. IPW estimates are also sensitive to misspecification of the π-model, because even mild lack of fit in outlying regions of the covariate space where $\pi_i \approx 0$ translates into large errors in the weights. These shortcomings of IPW estimators have been known for many years; see, for example, the comments of Little and Rubin (1987, Section 4.4.3) on the IPW methods for nonresponse proposed by Cassel, Särndal and Wretman (1983).

To see how well IPW performs on the artificial population described in Section 1.4, we created 1000 samples of $n = 200$ and $n = 1000$ units each and, for each sample, estimated the propensity scores in two ways. First, we fit a correctly specified π-model, regressing the t_i's on z_{i1}, z_{i2}, z_{i3} and z_{i4} using a logit link. Second, we fit the incorrect model which replaces the

z_{ij}'s with x_{ij}'s. The behavior of the four IPW estimates (3)–(4) under the correctly and incorrectly specified π-models is summarized in Table 1. In this table, "Bias" is the average difference between the estimate and $\mu = 210$, and "% Bias" is the bias as a percentage of the estimate's standard deviation. (A useful rule of thumb is that the performance of interval estimates and test statistics begins to deteriorate when the bias of the point estimate exceeds about 40% of its standard deviation.) "RMSE" is square root of the average value of $(\hat{\mu} - \mu)^2$. Examining Table 1, we see that the IPW estimates are biased when the π-model is misspecified, and the biases are accompanied by huge losses in precision. In fact, the bias and RMSE actually get worse as the sample size grows! IPW estimates have higher variance than other procedures examined in this article even when the propensities are correctly modeled, but when the π-model is incorrect, the method breaks down. Interestingly, NR weighting performs better than POP weighting; we do not know whether the superiority of NR is a peculiar feature of this example or if it tends to hold more generally.

The poor performance of IPW is due in part to occasional highly erratic estimates produced by a few enormous weights. In practice, a good data analyst would never use a simple IPW estimator if the weights were too extreme. Unusually large weights may be taken as a sign of model failure, prompting the researcher to revise the π-model. Outlying weights may be truncated or shrunk to more sensible values, or the offending units with large weights may be removed. (Removing these units is not recommended, because the respondents with the lowest propensities are in fact those that contain the best information for predicting nonrespondent behavior.) Analysts who apply IPW to real problems quickly learn that it often cannot be used without ad hoc modifications. The column of Table 1 labeled "MAE" reports the median of the absolute errors $|\hat{\mu} - \mu|$, which discards the worst 50% of the estimates. Even by this robust measure of precision, IPW performs more poorly than the other methods we examine when the π-model is misspecified, and it fails to improve as the sample size grows.

In many applications of IPW methodology, weights are obtained by logistic regression. Logistic models can be a poor way to estimate response propensities, because ML estimates from the logistic model are not resistant to outliers (Pregibon, 1982). A promising alternative is the robit model, which replaces the logistic link by the cumulative distribution function of a Student's t-distribution with ν degrees of freedom (Albert and Chib, 1993). The logit link is well approximated by the robit with $\nu = 7$, and smaller values of ν lead to estimates that are more robust (Liu, 2004). In this example, robit models produce minor improvements when the covariates are correct and major improvement when the covariates are wrong. We found that, with samples of $n = 1000$, replacing the logit link by robit with $\nu = 4$ reduces the bias and RMSE by nearly 50% when the π-model is incorrect. If IPW must be used, replacing the logistic regression with a more robust procedure can be advantageous.

2.2 Propensity-Based Stratification

To mitigate the dangers of extreme weights and misspecification of the π-model, some prefer to coarsen the estimated propensity scores into a few categories and compute weighted averages of the mean response across categories. In the context of survey nonresponse, the technique is known as weighting-cell estimation or adjustment (Oh and Scheuren, 1983; Little, 1986; Little and Rubin, 2002). Strong ignorability implies that y_i and t_i are conditionally independent within

TABLE 1
Performance of IPW estimators of μ over 1000 samples from the artificial population

Sample size	π-model	Method	Bias	% Bias	RMSE	MAE
(a) $n = 200$	Correct	*IPW-POP*	−0.27	−7.0	3.86	2.43
		IPW-NR	−0.29	−8.2	3.60	2.36
	Incorrect	*IPW-POP*	1.58	19.2	8.35	3.32
		IPW-NR	0.61	10.3	5.99	3.03
(b) $n = 1000$	Correct	*IPW-POP*	−0.01	−0.5	1.81	1.16
		IPW-NR	−0.03	−1.5	1.68	1.09
	Incorrect	*IPW-POP*	5.05	45.9	12.10	2.80
		IPW-NR	3.22	49.1	7.29	2.34

any subpopulation for which $\pi_i(x_i)$ is constant (Rosenbaum and Rubin, 1983). In classes of constant propensity, the mean values of y_i for respondents and nonrespondents are equal, which implies that

$$\mu = \int E(y_i \mid \pi_i, t_i = 1)\, dP(\pi_i). \tag{5}$$

Suppose we fit a π-model and define strata $s = 1, \ldots, S$ by grouping units whose $\hat{\pi}_i$'s are similar. Define $c_{is} = 1$ if unit i belongs to stratum s and 0 otherwise. The $\hat{\pi}$-stratified estimate of μ approximates (5) by a weighted average of respondents' mean in each stratum, weighted by the proportion of sample units in that stratum,

$$\hat{\mu}_{strat\text{-}\pi} = \sum_{s=1}^{S} \left(\frac{\sum_i c_{is}}{n} \right) \left(\frac{\sum_i c_{is} t_i y_i}{\sum_i c_{is} t_i} \right). \tag{6}$$

Similarly, a $\hat{\pi}$-stratified estimate of $\mu^{(0)}$ weights the respondents' mean in each stratum by the proportion of nonrespondents in that stratum,

$$\hat{\mu}^{(0)}_{strat\text{-}\pi} = \sum_{s=1}^{S} \left(\frac{\sum_i c_{is}(1 - t_i)}{n^{(0)}} \right) \left(\frac{\sum_i c_{is} t_i y_i}{\sum_i c_{is} t_i} \right),$$

which may be combined with $\bar{y}^{(1)}$ as in (4) to produce another estimate of $\hat{\mu}$. Rosenbaum and Rubin (1983) suggest classifying units into $S = 5$ strata defined by the sample quintiles of $\hat{\pi}_i$, as this tends to eliminate more than 90% of the selection bias if the π-model is correct (Cochran, 1968).

The performance of the $\hat{\pi}$-stratified estimator (6) over the 1000 samples from the artificial population is summarized in Table 2. Comparing these results to those of Table 1, we see that stratification is less effective than IPW at removing bias when the π-model is correct, but the stratified estimators still outperform IPW in terms of RMSE. The increase in bias incurred by coarsening the $\hat{\pi}_i$'s is easily offset by the greater efficiency that results from stabilizing the largest weights. When the π-model is misspecified, the stratified estimators are more biased than IPW for samples of $n = 200$ but less biased when $n = 1000$, because the bias of the stratified estimators does not worsen as n increases.

2.3 Regression Estimation

IPW and $\hat{\pi}$-stratified estimators pay no heed to relationships between the covariates and y_i. Regression estimators, on the other hand, model y_i from x_i directly and use this information to predict the missing values. Because strong ignorability implies that

$$E(y_i \mid x_i, t_i = 0) = E(y_i \mid x_i, t_i = 1) = E(y_i \mid x_i),$$

we can regress y_i on x_i among the respondents, apply the estimated regression function to predict y_i for the entire sample, and then average the predicted values to obtain an estimate of μ. Let

$$\hat{\beta} = \left(\sum_j t_j x_j x_j^T \right)^{-1} \left(\sum_j t_j x_j y_j \right)$$

denote the ordinary least-squares (OLS) coefficients from the regression of y_i on x_i among the respondents, and let $\hat{m}_i = x_i^T \hat{\beta}$. The OLS regression estimate for μ is

$$\hat{\mu}_{OLS} = \frac{1}{n} \sum_i \hat{m}_i. \tag{7}$$

This estimate is unbiased if the y-model is true, that is, if $E(y_i \mid x_i) = x_i^T \beta$ for some $\beta \in \mathcal{R}^p$; in addition, it is highly efficient if $\sigma_i^2 = V(y_i \mid x_i)$ is nearly constant. If the response is heteroscedastic, efficiency can be improved by replacing $\hat{\beta}$ with a weighted least-squares estimate with weights proportional to σ_i^{-2}.

From our 1000 samples, we computed OLS regression estimates under a correct y-model (regressing y_i on the z_{ij}'s) and an incorrect y-model (regressing y_i on the x_{ij}'s). The results, which are summarized in Table 3, verify that the bias is indeed removed when the y-model is correct but not when the model is wrong.

TABLE 2
Performance of propensity-stratified estimators over 1000 *samples from the artificial population*

Sample size	π-model	Method	Bias	% Bias	RMSE	MAE
(a) $n = 200$	Correct	*strat-π*	−1.15	−38.1	3.22	2.17
	Incorrect	*strat-π*	−2.82	−87.7	4.28	3.13
(b) $n = 1000$	Correct	*strat-π*	−1.08	−81.5	1.71	1.18
	Incorrect	*strat-π*	−2.87	−202.7	3.19	2.83

TABLE 3
Performance of ordinary least-squares regression estimators over 1000 *samples from the artificial population*

Sample size	y-model	Method	Bias	% Bias	RMSE	MAE
(a) $n = 200$	Correct	*OLS*	−0.08	−3.4	2.48	1.68
	Incorrect	*OLS*	−0.57	−17.7	3.26	2.24
(b) $n = 1000$	Correct	*OLS*	−0.00	−0.1	1.17	0.79
	Incorrect	*OLS*	−0.84	−56.0	1.72	1.15

Comparing the RMSE values in this table with those in Tables 1 and 2, we see that estimates based on the incorrect y-model are more stable and efficient than those based on the incorrect π-model. The bias that remains due to misspecification of the y-model is not large in absolute terms, but it is still troubling in samples of $n = 1000$ because there it amounts to more than 50% of a standard error and begins to impair tests and intervals. Difficulties with parametric missing-data methods arise when the uncertainty due to the model specification, which is rarely accounted for, grows relative to the sampling variation under the assumed model. In those cases, a point estimate based on a misspecified but reasonable y-model may still perform better than other estimates, but the analyst is too optimistic about its precision.

The performance of the regression estimate depends heavily on the strength of the correlation R between y_i and $\hat{m}_i$. As R^2 approaches 0, $\hat{\mu}_{OLS}$ converges to $\bar{y}^{(1)}$ and suffers from the same bias as this naive estimate. As R^2 approaches 1, it converges to the mean of the full sample, $\bar{y} = n^{-1} \sum_i y_i$, which is strongly robust. Therefore, if the y-model has strong prediction, the regression estimator tends to dominate other methods in terms of bias, efficiency and robustness. The IPW and π-stratified estimators, on the other hand, break down as the predictive ability of the π-model increases.

2.4 Stratification by Propensity Scores and Predicted Values

We saw that the efficiency and robustness of an IPW estimator can be enhanced by coarsening the $\hat{\pi}_i$'s into a small number of categories. The efficiency of these estimators can be further increased by adding covariates to the π-model that are predictive of y_i, even if these covariates are unrelated to t_i (Lunceford and Davidian, 2004). Vartivarian and Little (2002) show that further improvement is possible if we cross-classify units by estimated propensity scores and covariates that are strongly related to the outcome. The idea of combining estimated propensity scores with additional covariate information is not new. For example, Rosenbaum and Rubin (1985) recommended matching respondents to nonrespondents or vice versa by a Mahalanobis-distance criterion based on x_i within calipers defined by the estimated propensity scores. Numerous authors have performed regression adjustments based on x_i within strata defined by $\hat{\pi}$; for references, see D'Agostino (1998).

To illustrate a simple version of this idea, suppose that we create 25 strata by cross-classifying units into cells defined by the sample quintiles of $\hat{\pi}_i$ and $\hat{m}_i$, and then estimate μ by a weighted average of the mean response across the cells. If we apply this procedure over repeated samples, we will occasionally encounter a cell that contains nonrespondents but no respondents, in which case the stratified estimates are undefined. For those samples, we must modify the estimate in some fashion. For example, we may collapse adjacent rows or columns in the 5×5 table, fuse the offending cell with an adjacent cell, impute the missing cell mean by a regression estimate that assumes the response surface has row and column effects but no interactions, and so on.

In general, we dislike procedures that require frequent ad hoc adjustments unless their operating characteristics are well understood and the variance of the adaptive estimator can be approximated. Nevertheless, it is interesting to see how well the method performs in our artificial example. For each of our samples, we computed $(\hat{\pi} \times \hat{m})$-stratified estimators using all four possible combinations of correct and incorrect π- and y-models. Missing cell means were imputed by a regression procedure that assumes no row-by-column interactions. The results, which are summarized in Table 4, show that dual stratification produces a crude kind of double robustness; the bias is relatively low if the π-model is correctly specified or if the y-model is correctly specified. Under strong ignorability, a stratified estimate of the population mean will be unbiased if

TABLE 4
Performance of propensity and fitted-value stratified estimators over 1000 *samples from the artificial population*

Sample size	π-model	y-model	Method	Bias	% Bias	RMSE	MAE
(a) $n = 200$	Correct	Correct	*strat-πm*	−0.34	−11.6	2.92	1.90
		Incorrect	*strat-πm*	−0.59	−18.4	3.25	2.19
	Incorrect	Correct	*strat-πm*	−0.49	−17.1	2.89	1.96
		Incorrect	*strat-πm*	−2.00	−61.4	3.82	2.62
(b) $n = 1000$	Correct	Correct	*strat-πm*	−0.27	−21.1	1.31	0.87
		Incorrect	*strat-πm*	−0.45	−33.7	1.42	0.92
	Incorrect	Correct	*strat-πm*	−0.51	−39.6	1.38	0.92
		Incorrect	*strat-πm*	−2.10	−148.7	2.53	2.11

either the true π_i's or the true m_i's are constant within strata.

It is also useful to compare the results in Table 4 where both models are incorrect with those of Table 2 where the π-model is incorrect. This comparison shows that a π-stratified estimator can be improved with predicted values for the y_i's, even if those predictions come from a coarsened, misspecified y-model. The key idea of DR estimation—reducing your reliance on one model by specifying two—does produce modest gains in this example over estimates based on a π-model alone. Comparing Tables 3 and 4, however, we find that the approximate DR procedure based on two incorrect models performs worse than OLS based on the incorrect y-model. When neither model is exactly true, two models are not necessarily better than one.

3. CONSTRUCTING DOUBLY ROBUST ESTIMATES

3.1 Regression Estimation with Residual Bias Correction

Cassel, Särndal and Wretman (1976, 1977) introduced a family of "generalized regression estimators" for population means and totals that combine model-based predictions for y_i with inverse-probability weights. These methods, which are part of a methodology called model-assisted survey estimation (Särndal, Swensson and Wretman, 1992), are highly efficient when the y-model is true, yet remain asymptotically unbiased when the y-model is misspecified. In the original formulation, the response probabilities were a known feature of the survey design, but with uncontrolled nonresponse the propensities may be estimated under a π-model (Cassel, Särndal and Wretman, 1983).

To understand how these generalized regression estimators work, consider the simple regression estimator (7). If the regression model holds in the sense that $E(y_i|x_i) = x_i^T\beta$ for some $\beta \in \mathcal{R}^p$, then on average the predictions $\hat{m}_i = x_i^T\hat{\beta}$ will be neither too high nor too low; the mean of the estimated residuals $\hat{\varepsilon}_i = y_i - \hat{m}_i$ in the population will be zero. Of course, residuals are seen only for sampled respondents. We can, however, consistently estimate the mean residual for the full population if we have access to a π-model, and this estimate can in turn be used to correct the OLS estimate for bias arising from y-model failure. Cassel, Särndal and Wretman (1976) proposed the bias-corrected estimate

$$\hat{\mu}_{BC\text{-}OLS} = \hat{\mu}_{OLS} + \frac{1}{n}\sum_i t_i\hat{\pi}_i^{-1}\hat{\varepsilon}_i. \tag{8}$$

Notice that if the y-model is true, then $E(\hat{\varepsilon}_i) = 0$, and the second term on the right-hand side of (8) has expectation zero for arbitrary $\hat{\pi}_i$'s. If the π-model is true, then the second term consistently estimates (minus one times) the bias of the first term. Therefore, this estimate is doubly robust.

Many variations on this approach are possible. For example, we can normalize the weights in the correction term, so that the estimate becomes

$$\hat{\mu}_{OLS} + \frac{\sum_i t_i\hat{\pi}^{-1}\hat{\varepsilon}_i}{\sum_i t_i\hat{\pi}^{-1}}.$$

Or we can replace the POP weights with NR weights, so that the correction term estimates the mean residual in the population of nonrespondents. A bias-corrected estimate of $\mu^{(0)}$ based on this idea is

$$\frac{\sum_i (1-t_i)x_i^T\hat{\beta}}{\sum_i (1-t_i)} + \frac{\sum_i t_i\hat{\pi}^{-1}(1-\hat{\pi}_i)\hat{\varepsilon}_i}{\sum_i t_i\hat{\pi}^{-1}(1-\hat{\pi}_i)},$$

which can be combined with $\bar{y}^{(1)}$ as in (4) to produce another DR estimate for $\hat{\mu}$. A third possibility is to replace the weighted correction term with a $\hat{\pi}$-stratified estimate of $E(\hat{\varepsilon}_i)$. Using five strata based on sample quintiles of $\hat{\pi}_i$ would remove over 90% of the bias from the OLS estimate if the π-model were true and reduce problems of instability caused by very large weights.

A more general version of (8) was independently proposed by Robins, Rotnitzky and Zhao (1994) for estimating population-average regression coefficients from incomplete data. Suppose U_i is the contribution of sample unit i to a vector-valued quasi-score function for the regression of y_i on an arbitrary set of covariates. As noted in Section 2.1, the solution to $\sum_i w_i U_i = 0$ with $w_i = t_i \hat{\pi}_i^{-1}$ provides a consistent and asymptotically normal estimate of the population-average regression coefficients if the model used to estimate the π_i's is correct. Now suppose that we change the estimating equations to $\sum_i [w_i U_i + (1 - w_i)\phi_i] = 0$, where $\phi_i = \phi(x_i)$ is an arbitrary term that may depend on x_i but not on y_i. The mean of the additional term $(1 - w_i)\phi_i$ is essentially zero if the π-model is true, because $E(w_i) = E(E(w_i \mid x_i)) = E(\hat{\pi}_i^{-1}\pi_i) \approx 1$. Therefore, the solution to these augmented inverse-probability weighted (AIPW) estimating equations is again consistent and asymptotically normally distributed under a correct π-model. Robins and Rotnitzky (1995) demonstrate that a judicious choice for ϕ_i can greatly improve upon the efficiency of the simple IPW estimator. In particular, choosing $\phi_i = E(U_i \mid x_i)$, where the expectation is taken with respect to the distribution for y_i given x_i, produces a locally semiparametric efficient estimator, the most efficient estimator within this class. This estimate is DR, maintaining its consistency if either the π-model or y-model is correct (Scharfstein, Rotnitzky and Robins, 1999). Taking $U_i = (y_i - \mu)/\sigma^2$, and estimating $E(y_i \mid x_i)$ by $(\hat{m}_i - \mu)/\sigma^2$ where $\hat{m}_i = x_i^T \hat{\beta}$, the solution to the AIPW estimating equation reduces to the generalized regression estimator (8). More generally, it becomes the solution to

$$\frac{1}{n}\sum_i \hat{U}_i + \frac{1}{n}\sum_i t_i \hat{\pi}_i^{-1}(U_i - \hat{U}_i) = 0, \tag{9}$$

where $\hat{U}_i$ is the quasi-score function U_i with y_i replaced by $\hat{m}_i$.

By analogy to (8), (9) can be viewed as a predicted estimating equation with a residual bias correction. Any of the variations on (8) described above—normalizing the weights, switching to NR weights, or switching from an IPW bias correction to a $\hat{\pi}$-stratified one—can be applied to (9) as well. Modifications like these would take the estimator outside of the AIPW class. Nevertheless, these changes could potentially improve performance when either or both models are misspecified.

The performance of the bias-corrected regression estimate (8) in our simulated example is summarized in Table 5. The bias of this estimate does indeed vanish when either of the two models is correct. Moreover, comparing these results to those from the IPW-POP estimates in Table 1, we see that augmenting the IPW procedure by information from a correct y-model does indeed increase the efficiency. In the more realistic condition where both models are misspecified, however, this DR estimate does worse than IPW. A similar pattern emerges if we compare the new results to those from the simple OLS regression estimate in Table 3. Bias from an incorrect y-model is repaired by a π-model if the latter is correct. When both models are misspecified, however, the DR procedure is substantially worse than OLS. Like IPW, estimates from this DR procedure often behave erratically because one or more weights are occasionally enormous. Even if we

TABLE 5
Performance of bias-corrected regression estimators over 1000 *samples from the artificial population*

Sample size	π-model	y-model	Method	Bias	% Bias	RMSE	MAE
(a) $n = 200$	Correct	Correct	*BC-OLS*	−0.08	−3.4	2.48	1.68
		Incorrect	*BC-OLS*	0.25	7.5	3.28	2.17
	Incorrect	Correct	*BC-OLS*	−0.08	−3.3	2.48	1.70
		Incorrect	*BC-OLS*	−5.12	−43.0	12.96	3.54
(b) $n = 1000$	Correct	Correct	*BC-OLS*	0.00	−0.1	1.17	0.79
		Incorrect	*BC-OLS*	0.06	3.4	1.75	1.02
	Incorrect	Correct	*BC-OLS*	−0.02	−1.4	1.49	0.80
		Incorrect	*BC-OLS*	−21.03	−13.5	157.21	5.32

trim away these "bad" samples and judge the performance by the MAE, however, the new procedure is still worse than IPW and OLS. Once again, two models are not necessarily better than one.

Why does this DR estimate fail to perform better than IPW and OLS even though the π- and y-models are reasonably close to being true? The local semiparametric efficiency property, which guarantees that the solution to (9) is the best estimator within its class, was derived under the assumption that both models are correct. This estimate is indeed highly efficient when the π-model is true and the y-model is highly predictive. In our experience, however, if the π-model is misspecified, it is not difficult to find DR estimators that outperform this one by venturing outside the AIPW class. For this particular example, normalizing the POP weights, switching to NR weights, and using a $\hat{\pi}$-stratified bias correction all improve upon $\hat{\mu}_{BC\text{-}OLS}$. There are yet more ways to construct DR estimates, as we now describe.

3.2 Regression Estimation with Inverse-Propensity Weighted Coefficients

The correction term in $\hat{\mu}_{BC\text{-}OLS}$ repairs the bias in $\hat{\mu}_{OLS} = n^{-1}\sum_i x_i^T\hat{\beta}$ by estimating the mean residual in the full population. A different way to repair this bias is to move the estimated coefficients away from $\hat{\beta}$. Imagine that we could see the OLS coefficients based on the full sample,

$$\hat{\beta}_{SAMP} = \left(\sum_i x_i x_i^T\right)^{-1}\left(\sum_i x_i y_i\right).$$

A well-known property of OLS regression is that the sum of the estimated residuals $y_i - x_i^T\hat{\beta}_{SAMP}$, $i = 1,\ldots,n$, is zero if x_i includes a constant. This is an algebraic identity that holds regardless of the actual form of $E(y_i \mid x_i)$. This identity implies that the regression estimator based on $\hat{\beta}_{SAMP}$ replicates the mean of y_i in the full sample,

$$\frac{1}{n}\sum_i x_i^T\hat{\beta}_{SAMP} = \frac{1}{n}\sum_i y_i = \bar{y},$$

which is a strongly robust estimate of μ. We cannot compute $\hat{\beta}_{SAMP}$ from the observed data. But with propensities estimated from a π-model, we can compute a weighted least-squares (WLS) estimate

$$\hat{\beta}_{WLS} = \left(\sum_i t_i\hat{\pi}_i^{-1}x_i x_i^T\right)^{-1}\left(\sum_i t_i\hat{\pi}_i^{-1}x_i y_i\right).$$

In a well-behaved asymptotic sequence, $\hat{\beta}_{SAMP}$ and $\hat{\beta}_{WLS}$ both converge to the coefficients from the linear regression of y_i on x_i in the full population, regardless of whether that regression is an accurate portrayal of $E(y_i \mid x_i)$. If we compute a regression estimate for μ based on the WLS coefficients,

$$\hat{\mu}_{WLS} = \frac{1}{n}\sum_i x_i^T\hat{\beta}_{WLS}, \tag{10}$$

the difference between this estimate and $\bar{y}$ converges in probability to zero as $n \to \infty$, provided that the π-model holds.

From this discussion, we see that $\hat{\mu}_{WLS}$ consistently estimates μ if the π-model is true. If the y-model is true, then $\hat{\beta}_{WLS}$ will be an inefficient but consistent estimate of β, and $\hat{\mu}_{WLS}$ will again be consistent; thus it is DR.

In our simulated example, the WLS regression estimate is sometimes inferior to the bias-corrected OLS estimate (8) when one of the models is true, but much better when both models are misspecified (Table 6). Comparing $\hat{\mu}_{WLS}$ to $\hat{\mu}_{OLS}$, we see that the inverse-propensity weighted estimate of β effectively corrects

TABLE 6
Performance of regression estimators with inverse-propensity weighted coefficients over 1000 *samples from the artificial population*

Sample size	π-model	y-model	Method	Bias	% Bias	RMSE	MAE
(a) $n = 200$	Correct	Correct	*WLS*	−0.09	−3.4	2.48	1.68
		Incorrect	*WLS*	0.38	13.2	2.88	1.92
	Incorrect	Correct	*WLS*	−0.08	−3.4	2.48	1.68
		Incorrect	*WLS*	−2.20	−70.0	3.83	2.74
(b) $n = 1000$	Correct	Correct	*WLS*	0.00	−0.1	1.17	0.78
		Incorrect	*WLS*	0.16	12.0	1.35	0.92
	Incorrect	Correct	*WLS*	0.00	−0.1	1.17	0.78
		Incorrect	*WLS*	−2.99	−203.6	3.33	2.98

the bias from a wrong y-model if the π-model is correct, but makes matters worse if the π-model is wrong.

Once again, many variations on (10) are possible. Normalizing the POP weights $t_i\hat{\pi}_i^{-1}$ has no effect on $\hat{\beta}_{WLS}$, but one might consider switching to NR weights. Applying NR weights and averaging the predicted values of $x_i^T\beta$ among nonrespondents produces an estimate of $\hat{\mu}^{(0)}$, which can then be combined with $\bar{y}^{(1)}$ to produce another estimate of μ. Another possibility is to coarsen the weights into, say, five categories and compute a $\hat{\pi}$-stratified estimate of β.

3.3 Regression Estimation with Propensity-Based Covariates

A third general strategy for constructing a DR estimate is to incorporate functions of estimated propensities into the y-model as covariates. Ordinary regression estimates are based on the relationship

$$\mu = \int E(y_i \mid x_i)\, dP(x_i)$$

and achieve consistency if $E(y_i \mid x_i) = E(y_i \mid x_i, t_i = 1)$ is consistently estimated for all x_i. In practice, creating a model that gives unbiased predictions for y_i over the whole covariate space can be a daunting task, because real data often exhibit nonlinearities, interactions, etc. that are difficult to identify and portray, especially as the dimension of x_i grows. From (5), however, we see that requiring unbiased prediction for all x_i is much stronger than necessary; it would suffice to have unbiased prediction of $E(y_i \mid \pi(x_i)) = E(y_i \mid \pi(x_i), t_i = 1)$ for all $\pi(x_i) \in (0, 1)$. If we want to repair the bias in a parametric y-model, it makes sense to first identify and correct for lack of fit in the direction of the propensity score, because $\pi_i = \pi(x_i)$ is the coarsest summary of x_i that makes y_i and t_i conditionally independent (Rosenbaum and Rubin, 1983).

Consider the sample of $n = 200$ units from our artificial population that we examined in Section 1.4. Figure 4 shows the residuals $\hat{\varepsilon}_i$ from the linear regression of y_i on x_i, plotted against the linear predictors $\hat{\eta}_i$ from the logistic regression of t_i on x_i, for the $n^{(1)} = 100$ responding units. A least-squares line fit to this plot has an intercept and slope of zero, because the predictor is a perfect linear combination of covariates already in the y-model. A smooth curve created by a local polynomial (loess) fit, however, suggests that predictions from the y-model tend to be slightly low in the middle of the propensity scale and slightly high at the extremes. The bias could be corrected by fitting a generalized additive model (Hastie and Tibshirani, 1990) that allows the

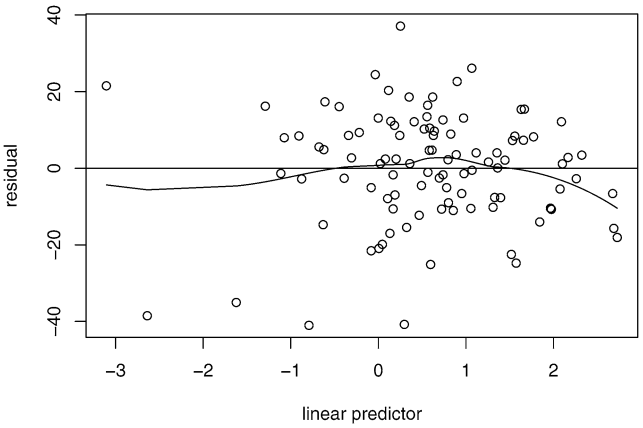


FIG. 4. *Scatterplot of raw residuals from y-model against linear predictors from a π-model, with least-squares and local polynomial fit.*

mean of y_i to vary smoothly with $\hat{\pi}_i$ in a nonparametric fashion. Little and An (2004) incorporated a smoothing spline hased on $\hat{\eta}_i$ and demonstrated that the resulting regression estimate of μ is DR in the following sense: If the y-model correctly describes $E(y_i \mid x_i)$ before the propensity-related terms are added, these additional terms (or any other functions of x_i) merely cause the model to be overfitted and add mean-zero noise to the predicted values. If the π-model is correct, then (5) guarantees consistent estimation of μ, as long as the mean of y_i varies smoothly with π_i and this relationship can be arbitrarily well approximated by the linear combination of basis functions added to the model. In the latter case, the propensity-related covariates completely remove the bias for estimating μ, and any additional information provided by x_i merely serves to make the estimate more precise.

In the spirit of Little and An (2004), let $S_i = S(\hat{\eta}_i)$ denote a vector of basis functions (e.g., a spline basis) that can serve to approximate the relationship between the mean of y_i and $\hat{\eta}_i$, the estimated linear predictor from the π-model. Let $x_i^* = (x_i^T, S_i^T)^T$ denote the augmented vector of covariates, and let

$$\hat{\beta}^* = \left(\sum_i t_i x_i^* x_i^{*T}\right)^{-1} \left(\sum_i t_i x_i^* y_i\right) \tag{11}$$

denote the OLS-estimated coefficients from the augmented y-model. If S_i is a spline basis of degree $k \geq 1$, the matrix inverse in (11) will not exist, because S_i will contain a constant and a linear function of $\hat{\eta}_i$, which are themselves linear functions of x_i. Problems of collinearity can be alleviated by switching to a generalized inverse or by removing the offending terms

TABLE 7
Performance of propensity-covariate (four dummy indicators) regression estimators over 1000 *samples from the artificial population*

Sample size	π-model	y-model	Method	Bias	% Bias	RMSE	MAE
(a) $n = 200$	Correct	Correct	π-*cov*	−0.09	−3.4	2.48	1.69
		Incorrect	π-*cov*	−0.39	−13.5	2.93	2.00
	Incorrect	Correct	π-*cov*	−0.09	−3.4	2.48	1.68
		Incorrect	π-*cov*	−1.27	−38.6	3.51	2.43
(b) $n = 1000$	Correct	Correct	π-*cov*	0.00	−0.1	1.17	0.79
		Incorrect	π-*cov*	−0.55	−42.1	1.41	0.87
	Incorrect	Correct	π-*cov*	0.00	−0.2	1.17	0.79
		Incorrect	π-*cov*	−1.49	−100.6	2.10	1.56

from S_i; either approach leads to the same predicted values $\hat{m}_i^* = x_i^{*T}\hat{\beta}^*$. The propensity-covariate regression estimate for μ is then

$$\hat{\mu}_{\pi\text{-}cov} = \frac{1}{n}\sum_i \hat{m}_i^*. \tag{12}$$

A particular case of (12) was proposed by Scharfstein, Rotnitzky and Robins (1999) who took $S_i = \hat{\pi}_i^{-1}$. Using the inverse-propensity as a single additional covariate is sufficient to achieve double robustness, because the estimate then becomes the solution to an AIPW estimating equation (Bang and Robins, 2005).

We tried (12) in our simulated example with a variety of spline bases: a quadratic spline with a single knot at the median of $\hat{\eta}_i$, a linear spline with knots at the quartiles, and so on. We found that these polynomial splines occasionally produced erratic predicted values of y_i for nonrespondents with low propensities, driving up the variance of the regression estimate. The best performance was achieved by simply coarsening the $\hat{\eta}_i$'s into five categories and creating four dummy indicators to distinguish among them. In other words, we approximated the relationship between the mean of y_i and $\hat{\pi}_i$ by a piecewise-constant function with discontinuities at the sample quintiles of $\hat{\pi}_i$. The performance of this estimate is summarized in Table 7. It performs better than any of the other DR methods when the π-model and y-model are both incorrect. It performs better than any method based on a π-model alone. Yet it is still inferior to the simple OLS regression estimate under the incorrect y-model.

In contrast, the regression estimate of Scharfstein, Rotnitzky and Robins (1999) that uses the inverse-propensity as a covariate behaves poorly under a misspecified π-model (Table 8). The performance of this method is disastrous when some of the estimated propensities are small.

4. DISCUSSION

Double robustness is an interesting theoretical property that can arise in many ways. It does not necessarily translate into good performance when neither

TABLE 8
Performance of inverse-propensity covariate regression estimators over 1000 *samples from the artificial population*

Sample size	π-model	y-model	Method	Bias	% Bias	RMSE	MAE
(a) $n = 200$	Correct	Correct	$1/\pi$-*cov*	−0.09	−3.7	2.48	1.69
		Incorrect	$1/\pi$-*cov*	1.66	37.6	4.70	2.84
	Incorrect	Correct	$1/\pi$-*cov*	56.5	3.1	1804	1.76
		Incorrect	$1/\pi$-*cov*	−4236	−3.4	1.3×10^5	7.79
(b) $n = 1000$	Correct	Correct	$1/\pi$-*cov*	0.00	−0.1	1.17	0.78
		Incorrect	$1/\pi$-*cov*	0.59	31.3	1.97	1.26
	Incorrect	Correct	$1/\pi$-*cov*	−50.4	−3.2	1593	0.83
		Incorrect	$1/\pi$-*cov*	−7527	−3.3	2.3×10^5	5.75

model is correctly specified. Some DR estimators have been known to survey statisticians since the late 1970s. In survey contexts, these methods are not thought of as doubly robust but simply as robust, because the propensities are a known feature of the sample design. When propensities are unknown and must be estimated, care should be taken to select an estimator that is not overly sensitive to misspecification of the propensity model.

No single example can effectively represent all missing-data problems that researchers will see in practice. We constructed our simulation to vaguely resemble a quasi-experiment to measure the effect of dieting on body mass index (BMI) in a large sample of high-school students. The study has a pre-post design. Covariates x_i measured at baseline include demographic variables, BMI, self-perceived weight and physical fitness, social acceptance and personality measures. The treatment t_i is dieting (0 = yes, 1 = no) and the outcome y_i is BMI one year later. The goal is to estimate an average causal effect of dieting among those who actually dieted. For that purpose, it suffices to treat the dieters as nonrespondents, set their BMI values to missing, and apply a missing-data method to estimate what the mean BMI for this group would have been had they not dieted. A simple linear regression of y_i on x_i among the nondieters yielded an R^2 of 0.81, just as in our simulated data, and boxplots of the linear predictors from a logistic propensity model looked similar to those of Figure 3(e). The large degree of overlap in the distributions of estimated propensities for the two groups makes a causal analysis seem feasible. The estimated propensities for some nondieters are very small. Keeping these nondieters in the analysis is highly desirable, because their covariate values closely resemble those of dieters; they provide excellent proxy information for predicting the missing values. Yet we found no obvious way to use these cases in an inverse-propensity weighted estimate, because they exerted too much influence.

Our simulation represents a situation where selection bias is moderate, good predictors of y_i are available, both models are approximately but not exactly true, and some estimated propensities are nearly zero. In situations like these, a DR procedure that does not rely on inverse-propensity weighting may perform reasonably well, but there is no guarantee that it will outperform a procedure based solely on a y-model. A model that predicts y_i reasonably well can enhance the performance of an approximate π-model. But in our simulations, we found no way to use the fitted propensities from the approximate π-model to reduce the bias from the approximate y-model. In every case, the bias correction applied to $\hat{\mu}_{OLS}$ tended to move the estimate in the wrong direction.

Some might argue that inference about $E(y_i)$ in the full population should not be attempted in a situation like the one shown in Figure 3(e), where the estimated propensities for a few nonrespondents fall outside the range of those seen among the respondents. In our opinion, requiring the empirical support of $P(\hat{\pi}_i \mid t_i = 1)$ to completely cover that of $P(\hat{\pi}_i \mid t_i = 0)$ is too stringent, especially given the sensitivity of these ranges to minor changes in the π-model. In all 1000 samples of $n = 200$ and $n = 1000$, the distributions of the estimated propensities overlapped sufficiently to compute a stratified estimate with five quintile-based groups. Moreover, although some of the estimated propensities were very small, none of the true propensities were actually zero, so the conditions required for DR estimation were not violated. Small propensities frequently occur when missing data are not missing by design, and data analysts need guidance on how to deal with them. One who would argue for the routine use of any kind of inverse-propensity weighted estimator ignores the obvious fact that these estimators cannot be used routinely.

Other evaluations of DR methods under dual misspecification have yielded mixed results, because the nature of the problems and degree of misspecification have varied. Davidian, Tsiatis and Leon (2005) presented a DR procedure analogous to $\hat{\mu}_{BC\text{-}OLS}$ for a pre–post analysis of a randomized clinical trial with dropout. Schafer and Kang (2005) evaluated their procedure and found that it performed slightly better than a parametric method based on multiple imputation (Rubin, 1987). In that analysis, each of the two treatment groups had its own y-model with R^2's of about 0.5. Each group also had its own π-model, and the smallest fitted propensities were 0.14 and 0.07. It thus appears that the use of AIPW estimating equations can produce modest gains over a parametric method when the predictive power of the y-model is not too strong and the estimated propensities do not get close to zero.

The only other simulations that we know of that compare the performance of DR and non-DR methods under dual misspecification are those of Bang and Robins (2005). Their first example pertains to the estimation of a population mean μ. They compare the performance of three methods—the unnormalized IPW estimate $n^{-1}\sum_i t_i \hat{\pi}_i^{-1} y_i$, the ordinary regression estimate $n^{-1}\sum_i x_i^T \hat{\beta}$, and the propensity-covariate estimate (12) that augments the y-model with $1/\hat{\pi}_i$—when

the π- and y-models are correct and incorrect. In that example, the predictive ability of the correct y-model among the respondents is very strong ($R^2 = 0.94$), but the incorrect version is worthless ($R^2 = 0.001$); thus it maximally punishes the simple regression method when the y-model is misspecified. This approximates a situation where an analyst wants to impute missing values but has no idea how to use the covariates, so he simply ignores them and replaces all the missing values with $\bar{y}_i^{(1)}$. It may also represent a situation where all of the confounders are hidden from the analyst for purposes of y-modeling (though not, strangely, for purposes of π-modeling). Another noteworthy feature of that example is that all of the covariates in the propensity model have been dichotomized and the selection mechanism is weak; when n is large, all of the estimated propensities fall nicely between 0.25 and 0.5. Bang and Robins (2005) present three additional examples to demonstrate the performance of DR estimates in more elaborate problems. In each case, all of the predictors in their π-models were dichotomized, which helps to keep the estimated propensities away from zero. Despite these features, none of their examples supports the claim that a DR method based on two incorrect models is clearly and simultaneously better than an IPW procedure based on an incorrect π-model and a simple imputation procedure based on an incorrect y-model. Only in the fourth example did the DR method outperform both of its competitors under dual misspecification, and even there the advantage was slight. Thus the results of Bang and Robins (2005) support our contention that two wrong models are not necessarily better than one.

ACKNOWLEDGMENT

This research was supported by National Institute on Drug Abuse Grant P50-DA10075.

REFERENCES

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189

BINDER, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Internat. Statist. Rev.* **51** 279–292. MR0731144

CARPENTER, J., KENWARD, M. and VANSTEELANDT, S. (2006). A comparison of multiple imputation and inverse probability weighting for analyses with missing data. *J. Roy. Statist. Soc. Ser. A* **169** 571–584.

CASSEL, C. M., SÄRNDAL, C. E. and WRETMAN, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63** 615–620. MR0445666

CASSEL, C. M., SÄRNDAL, C. E. and WRETMAN, J. H. (1977). *Foundations of Inference in Survey Sampling*. Wiley, New York. MR0652527

CASSEL, C. M., SÄRNDAL, C. E. and WRETMAN, J. H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete Data in Sample Surveys* **III**. *Symposium on Incomplete Data*, *Proceedings* (W. G. Madow and I. Olkin, eds.). Academic Press, New York.

COCHRAN, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24** 205–213. MR0228136

D'AGOSTINO, R. B. JR. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* **17** 2265–2281.

DAVIDIAN, M., TSIATIS, A. A. and LEON, S. (2005). Semiparametric estimation of treatment effect in a pretest–posttest study without missing data. *Statist. Sci.* **20** 261–301. MR2189002

GELMAN, A. and MENG, X. L., eds. (2004). *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. Wiley, New York. MR2134796

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall, London. MR2027492

GEWEKE, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57** 1317–1339. MR1035115

HAMMERSLEY, J. M. and HANDSCOMB, D. C. (1964). *Monte Carlo Methods*. Methuen, London. MR0223065

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London. MR1082147

HINKLEY, D. (1985). Transformation diagnostics for linear models. *Biometrika* **72** 487–496. MR0817563

HIRANO, K. and IMBENS, G. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catherization. *Health Services and Outcome Research Methodology* **2** 259–278.

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460

LITTLE, R. J. A. and AN, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statist. Sinica* **14** 949–968. MR2089342

LITTLE, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *Internat. Statist. Rev.* **54** 139–157.

LITTLE, R. J. A. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York. MR0890519

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, New York. MR1925014

LIU, C. (2004). Robit regression: A simple robust alternative to logistic and probit regression. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (A. Gelman and X. L. Meng, eds.) 227–238. Wiley, New York. MR2138259

LUNCEFORD, J. K. and DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* **23** 2937–2960.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London. MR0727836

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments: Essays on principles, Section 9. Translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. *Statist. Sci.* **5** (1990) 465–480. MR1092986

OH, H. L. and SCHEUREN, F. S. (1983). Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys* **II**. *Theory and Annotated Bibliography* (W. G. Madow, I. Olkin and D. B. Rubin, eds.) 143–184. Academic Press, New York.

PREGIBON, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics* **38** 485–498.

ROBINS, J. M. and ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* **90** 122–129. MR1325119

ROBINS, J. M. and ROTNITZKY, A. (2001). Comment on "Inference for semiparametric models: some questions and an answer," by P. J. Bickel and J. Kwon. *Statist. Sinica* **11** 920–936. MR1867326

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90** 106–121. MR1325118

ROTNITZKY, A., ROBINS, J. M. and SCHARFSTEIN, D. O. (1998). Semiparametric regression for repeated outcomes with ignorable nonresponse. *J. Amer. Statist. Assoc.* **93** 1321–1339. MR1666631

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. Springer, New York. MR1899138

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician* **39** 33–38.

RUBIN, D. B. (1974a). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educational Psychology* **66** 688–701.

RUBIN, D. B. (1974b). Characterizing the estimation of parameters in incomplete data problems. *J. Amer. Statist. Assoc.* **69** 467–474.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519

RUBIN, D. B. (2005). Causal inference using potential outcomes: design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331. MR2166071

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of a finite population total. *Biometrika* **76** 527–537. MR1040646

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer, New York. MR1140409

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London. MR1692799

SCHAFER, J. L. and KANG, J. D. Y. (2005). Discussion of "Semiparametric estimation of treatment effect in a pretest–postest study with missing data" by M. Davidian et al. *Statist. Sci.* **20** 292–295. MR2189002

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1096–1120 (with rejoinder 1135–1146). MR1731478

VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer, New York. MR1958123

VARTIVARIAN, S. and LITTLE, R. J. A. (2002). On the formation of weighting adjustment cells for unit nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association*. Amer. Statist. Assoc., Alexandria, VA.

WINSHIP, C. and SOBEL, M. E. (2004). Causal inference in sociological studies. In *Handbook of Data Analysis* (M. Hardy, ed.) 481–503. Thousand Oaks, Sage, CA.

Statistical Science
2007, Vol. 22, No. 4, 540–543
DOI: 10.1214/07-STS227C
Main article DOI: 10.1214/07-STS227
© Institute of Mathematical Statistics, 2007

Comment: Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data

Greg Ridgeway and Daniel F. McCaffrey

This article is an excellent introduction to doubly robust methods and we congratulate the authors for their thoroughness in bringing together the wide array of methods from different traditions that all share the property of being doubly robust.

Statisticians at RAND have been making extensive use of propensity score weighting in education (McCaffrey and Hamilton, 2007), policing and criminal justice (Ridgeway, 2006), drug treatment evaluation (Morral et al., 2006), and military workforce issues (Harrell, Lim, Castaneda and Golinelli, 2004). More recently, we have been adopting doubly robust (DR) methods in these applications believing that we could achieve further bias and variance reduction. Initially, this article made us second-guess our decision. The apparently strong performance of OLS and the authors' finding that no method outperformed OLS ran counter to our intuition and experience with propensity score weighting and DR estimators. We posited two potential explanations for this. First, we suspected that the high variance reported by the authors when using propensity score weights could result from their use of standard logistic regression. Second, stronger interaction effects in the outcome regression model might favor the DR approach.

1. METHODS

We felt the authors were somewhat narrow in their discussion of weighting by focusing only on propensity scores estimated by logistic regression in their simulation. The high variability in the weights reported by the authors could result from using this method. The authors state that none of the various IPW methods could overcome the problems with estimated propensity scores near 0 and 1, yet we believed that this is indicative of a problem with the propensity score estimator rather than IPW methods. In our experience weights estimated using a generalized boosted model (GBM) following the methods of McCaffrey, Ridgeway and Morral (2004) as implemented in the Toolkit for Weighting and Analysis of Nonequivalent Groups, the `twang` package for R, tend not to show the extreme behavior that resulted from logistic regression (Ridgeway, McCaffrey and Morral, 2006).

GBM is a general, automated, data-adaptive algorithm that can be used with a large number of covariates to fit a nonlinear surface and estimate propensity scores. GBM uses a linear combination of a large collection of piecewise constant basis functions to construct a regression model for dichotomous outcomes. Shrunken coefficients prevent the model from overfitting. The use of piecewise constants has the effect of keeping the estimated propensity scores relatively flat at the edges of the range of the predictors, yet it still produces well-calibrated probability estimates. This reduces the risk of the spurious predicted probabilities near 0 and 1 that cause problems for propensity score weighting. Many variants of boosting have appeared in machine learning and statistics literature and Hastie, Tibshirani and Friedman (2001) provide an overview. We optimized the number of terms in the GBM model to provide the best "balance" between the weighted covariate distributions $f(\mathbf{x}|t=1)$ and $f(\mathbf{x}|t=0)$. This approach to fitting propensity scores is fully implemented in the `twang` package.

We tested our conjectures about the performance of IPW and DR estimators based on GBM and in the presence of omitted interactions terms through a simulation experiment using the same design that the authors used.

Greg Ridgeway is Senior Statistician and Associate Director of the Safety and Justice Program at the RAND Corporation, Santa Monica, California 90407-2138, USA (e-mail: gregr@rand.org). Daniel F. McCaffrey is Senior Statistician and Head of the Statistics Group at the RAND Corporation, Pittsburgh, Pennsylvania 15213, USA (e-mail: danielm@rand.org).

Using their model from Section 1.4, we generated 1000 datasets and calculated the population and nonresponse estimator as they did for Tables 1, 3, 5 and 6. In addition, for each dataset we also estimated propensity scores using GBM (the `ps()` function optimizes the number of basis functions in the model to minimize the largest of the marginal Kolmogorov–Smirnov statistics). While their simulations do not test this, Kang and Schafer noted that choices other than logistic regression may be preferable and offered robit regression as a possibility. We included robit(1) propensity score estimates in our simulations as well.

In addition to experimenting with other propensity score estimators, we also expanded the simulation to add an interaction term equal to $20Z_1Z_2$ to the mean function for Y. R code for the simulation experiments is available upon request.

2. RESULTS

Table 1 shows the results for the IPW methods. The rows of the table correspond to the different estimators presented by Kang and Schafer. The row labeled "Model for Y" denotes whether the outcome OLS model used the Z variables or the "mis-transformed" X variables in the estimation or, in the case of the interaction experiments, whether the fitted model includes a Z_1Z_2 term. The 12 IPW estimators in Table 1 vary by the propensity score model (logistic, GBM or robit regression), the use of Z or X as covariates in the propensity score model, and the use of either population weighting (IPW-POP) or the nonresponder reweighting (IPW-NR). The elements of the table contain the ratio of the RMSE of the alternative estimators to the RMSE of OLS fit with the covariates listed in each column heading.

First note that IPW estimators with logistic regression using the X covariates have by far the largest RMSEs in the table. Second, while OLS seems to be preferable over IPW methods in the case where there is in truth no interaction, when the OLS models exclude an important interaction the IPW methods are preferable. When faced with the choice between OLS and IPW, the analyst must decide whether to hedge against an interaction and use IPW or choose OLS, hoping that the outcome model is specified correctly and consequently gaining a 60% improvement over GBM-based IPW or a 10% improvement over robit-based IPW.

The aim of DR estimators is to avoid this dilemma and the associated hedging by combining the benefits of both the outcome and selection models. Kang and Schafer's results suggest that current DR estimators can disappoint us. They show DR estimators having twice the RMSE as OLS estimators when both the outcome and selection models use the X covariates. We

TABLE 1
Simulation study results for IPW methods

Generated data:			**K&S model**		**K&S model with interaction**		
Model for Y:			**Fit with Z**	**Fit with X**	**Fit with Z and interactions**	**Fit with Z, no interactions**	**Fit with X**
OLS			**1.0 (1.16)**	**1.0 (1.64)**	**1.0 (1.35)**	**1.0 (3.58)**	**1.0 (5.00)**
Logistic	Z	IPW-POP	1.4	1.0	2.0	0.7	0.5
		IPW-NR	1.3	0.9	1.9	0.7	0.5
	X	IPW-POP	9.9	7.0	9.7	3.6	2.6
		IPW-NR	6.0	4.3	5.9	2.2	1.6
GBM	Z	IPW-POP	1.9	1.3	2.2	0.8	0.6
		IPW-NR	1.5	1.0	2.1	0.8	0.6
	X	IPW-POP	2.6	1.9	3.1	1.2	0.8
		IPW-NR	2.2	1.6	2.7	1.0	0.7
Robit	Z	IPW-POP	1.4	1.0	1.7	0.6	0.4
		IPW-NR	1.3	0.9	2.4	0.9	0.6
	X	IPW-POP	1.6	1.1	2.8	1.0	0.7
		IPW-NR	1.6	1.1	2.9	1.1	0.8

The rows define the model used for the propensity score weights and the columns define the variables used in the outcome regression. The cells show the ratio of the RMSE of the estimator to the RMSE of the OLS model that used the covariates listed in the column title. The actual RMSE of the OLS model is shown in parentheses.

TABLE 2
Simulation study results for DR methods

Generated data:			**K&S model**		**K&S model with interaction**		
Model for Y:			**Fit with Z**	**Fit with X**	**Fit with Z and interactions**	**Fit with Z, no interactions**	**Fit with X**
OLS			**1.0 (1.16)**	**1.0 (1.64)**	**1.0 (1.35)**	**1.0 (3.58)**	**1.0 (5.00)**
Logistic	Z	BC	1.0	1.0	1.0	0.6	0.4
		WLS	1.0	0.8	1.0	0.5	0.4
	X	BC	1.0	51.3	2.6	85.8	139.2
		WLS	1.0	2.0	1.0	1.1	1.2
GBM	X	BC	1.0	0.9	1.0	*0.6	0.6
		WLS	1.0	0.9	1.0	*0.5	0.6
Robit	X	BC	1.0	1.9	1.0	*0.5	1.2
		WLS	1.0	1.5	1.0	*0.5	1.0

*These estimators use Z in the propensity score model.
The rows define the model used for the propensity score weights and the columns define the variables used in the outcome regression. The cells show the ratio of the RMSE of the estimator to the RMSE of the OLS model that used the covariates listed in the column title. All GBM and robit models were fit using X with the exception of the "Fit with Z, no interactions" column for which they were fit with Z. The actual RMSE of the OLS model is shown in parentheses.

investigated this using the same propensity score models described previously and the bias corrected (BC) and weighted least squares (WLS) described by Kang and Schafer. Table 2 compares the relative efficiency of DR estimators in terms of the RMSE of the DR estimators compared to OLS. The most interesting comparisons are those for which both the propensity score model and the outcome regression model use X. Other combinations, such as the propensity score fit with X and the outcome regression fit with Z, are not realistic but are included for completeness.

The results clearly show that WLS with GBM dominates OLS. When the model for Y is correct, WLS is essentially as efficient as the OLS estimator. When the model for Y is incorrect, WLS with GBM can be significantly more efficient than OLS. GBM also outperforms the robit regression model that the authors suggested as an option. These results suggest that DR estimators might be reliable methods of buying insurance against model misspecification without paying a high price in lost efficiency.

3. SUMMARY

In the simulation the doubly robust estimators are particularly useful when the model is missing an important interaction between pretreatment variables. Exploratory data analysis could be used to find such missing terms in the model and hence the advantages of WLS might appear overstated. However, such exploratory analyses require modeling the outcome and present the opportunity for the model selection to be corrupted by the impact of alternative models on the estimated treatment effect. That is, the model might be chosen because it yields significant treatment effects. This type of model fitting removes one of the benefits of the propensity score approach, which is the ability to control for pretreatment variable prior to seeing the outcome to avoid the temptation or even the appearance of data snooping.

Doubly robust estimators with GBM appear to have the desired properties in this simulation study. When the model for the mean is correct, there is no cost for using the doubly robust estimator (bias corrected or WLS). They are essentially as efficient as the correctly specified OLS model. When the OLS model is incorrect, again the doubly robust estimators are at least as efficient as OLS and substantially more efficient when the OLS model is missing important interaction terms. While it is clear that more work on these estimators is needed, our results do suggest that doubly robust estimation should not be dismissed too quickly.

REFERENCES

HARRELL, M. C., LIM, N., CASTANEDA, L. and GOLINELLI, D. (2004). Working around the military: Challenges to military spouse employment and education. RAND Monograph MG-196-OSD, Santa Monica, CA.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. H. (2001). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*. Springer, New York. MR1851606

MCCAFFREY, D. F. and HAMILTON, L. S. (2007). Value-added assessment in practice: Lessons from the Pennsylvania value-added assessment system pilot project. RAND Technical Report TR-506, Santa Monica, CA.

MCCAFFREY, D. F., RIDGEWAY, G. and MORRAL, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Methods* **9** 403–425.

MORRAL, A. R., MCCAFFREY, D. F., RIDGEWAY, G., MUKHERJI, A. and BEIGHLEY, C. (2006). The relative effectiveness of 10 adolescent substance abuse treatment programs in the United States. RAND Technical Report TR-346, Santa Monica, CA.

RIDGEWAY, G. (2006). Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. *J. Quant. Crimінol.* **22** 1–29.

RIDGEWAY, G., MCCAFFREY, D. F. and MORRAL, A. R. (2006). Twang: Toolkit for weighting and analysis of nonequivalent groups. (Software and reference manual.) Available at http://cran.r-project.org/src/contrib/Descriptions/twang.html.

Statistical Science
2007, Vol. 22, No. 4, 544–559
DOI: 10.1214/07-STS227D
Main article DOI: 10.1214/07-STS227
© Institute of Mathematical Statistics, 2007

Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable

James Robins, Mariela Sued, Quanhong Lei-Gomez and Andrea Rotnitzky

1. GENERAL CONSIDERATIONS

We thank the editor Ed George for the opportunity to discuss the paper by Kang and Schaeffer.

The authors' paper provides a review of double-robust (equivalently, double-protected) estimators of (i) the mean $\mu = E(Y)$ of a response Y when Y is missing at random (MAR) (but not completely at random) and of (ii) the average treatment effect in an observational study under the assumption of strong ignorability. In our discussion we will depart from the notation in Kang and Schaeffer (throughout, K&S) and use capital letters to denote random variables and lowercase letter to denote their possible values.

In the missing-data setting (i), one observes n i.i.d. copies of $O = (T, X, TY)$, where X is a vector of always observed covariates and T is the indicator that the response Y is observed. An estimator of μ is double-robust (throughout, DR) if it remains consistent and asymptotically normal (throughout, CAN) when either (but not necessarily both) a model for the propensity score $\pi(X) \equiv P(T = 1|X) = P(T = 1|X, Y)$ or a model for the conditional mean $m(X) \equiv E(Y|X) = E(Y|X, T = 1)$ is correctly specified, where the equalities follow from the MAR assumption. The authors demonstrate, via simulation, that when a linear logistic model for the propensity score and a linear model for the mean of Y given X are both moderately misspecified, there exists a joint distribution under which the OLS regression estimator $\hat{\mu}_{OLS}$ of μ outperforms all candidate estimators that depend on a linear logistic maximum likelihood estimate of the propensity score, including all the DR estimators considered by the authors.

Near the end of their Section 1, the authors state that their simulation example "appears to be precisely the type of situation for which the DR estimators of Robins et al. were developed." They then suggest that their simulation results imply that the cited quotation from Bang and Robins (2005) is incorrect or, at the very least, misguided. We disagree with both the authors' statement and suggestion. First, the cited quote neither claims nor implies that when a linear logistic model for the propensity score and a linear model for the mean of Y given X are moderately misspecified, DR estimators always outperform estimators—such as regression, maximum likelihood, or parametric (multiple) imputation estimators—that do not depend on the estimated propensity score. Indeed, Robins and Wang (2000) in their paper "Inference for Imputation Estimators" stated the following:

> If nonresponse is ignorable, a locally semiparametric efficient estimator is doubly protected; i.e., it is consistent if either a model for nonresponse or a parametric model for the complete data can be correctly specified. On the other hand, consistency of a parametric multiple imputation estimator requires correct specification of a parametric model for the complete data. However, in cases in which the variance of the 'inverse probability' weights is very large, the sampling distribution of a locally semiparametric efficient (augmented inverse probability of response weighted) estimator can be markedly skew and highly variable, and a parametric imputation estimator may be preferred.

The just-quoted cautionary message of Robins and Wang (2000) is not far from K&S's take-home mes-

James Robins is Professor, Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA (e-mail: akaris@hsph.harvard.edu). Mariela Sued is Assistant Professor, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Argentina. Quanhong Lei-Gomez is a Graduate Student, Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA. Andrea Rotnitzky is Professor, Department of Economics, Di Tella University, Buenos Aires, Argentina.

sage. In Section 5 we show that, in the authors' simulation example, the variance of the estimated "inverse probability" weights is very large and the sampling distribution of their candidate DR estimators is skewed and highly variable. It follows that their example is far from the settings Bang and Robins had in mind when recommending the "routine use of DR estimators." Rather, their example falls squarely into the class for which Robins and Wang (2000) cautioned that a parametric imputation estimator may be preferable to DR estimators.

Even prior to Robins and Wang (2000), Robins, Rotnitzky and colleagues had published extensive warnings about, and simulation studies of, the hazards of highly variable "inverse probability" weights (Robins, Rotnitzky and Zhao, 1995, pages 113–115; Scharfstein, Rotnitzky and Robins, 1999, pages 1108–1113), although not specifically for DR estimators. Due to the fact that the problem of highly variable weights was not the focus of their paper and had already been discussed extensively in earlier papers by Robins and colleagues, Bang and Robins (2005) did not repeat Robins and Wang's (2000) cautionary message. In retrospect, had they done so or had the authors been aware of the Robins and Wang article, a misunderstanding could perhaps have been averted.

Whenever the "inverse probability" weights are highly variable, as in K&S's simulation experiment, a small subset of the sample will have extremely large weights relative to the remainder of the sample. In this setting, no estimator of the marginal mean $\mu = E(Y)$ can be guaranteed to perform well. That is why, in such settings, some "argue that inference about the mean $E(Y)$ in the full population should not be attempted," to quote from the authors' discussion. Yet, surprisingly, in the authors' simulation experiment, the regression estimator $\hat{\mu}_{OLS}$ performed very well with a mean squared error (MSE) less than any of their candidate DR estimators, all of which estimated the propensity score by maximum likelihood under a linear logistic model. The explanation is that, whether due to unusual luck or to "cherry-picking," the chosen data-generating distribution was as if optimized to have $\hat{\mu}_{OLS}$ perform well. Indeed, in Section 5, we "deconstruct" the chosen distribution and show that it possesses a number of specific, some rather unusual, features that together served to insure $\hat{\mu}_{OLS}$ would perform well even under K&S's misspecified models.

Now, even were the chosen joint distribution of (Y, T, X) optimized to have $\hat{\mu}_{OLS}$ perform extremely well as an estimator of $E(Y)$ on data (TY, T, X) in which Y is observed only when $T = 1$, such optimization would not guarantee that $\hat{\mu}_{OLS}$ would also perform well on the data $((1 - T)Y, T, X)$ in which Y is observed only when $T = 0$. Based on this insight, in Section 5, we repeat K&S's simulation study, except based on data $((1 - T)Y, T, X)$ rather than data (TY, T, X), and show that, indeed, $\hat{\mu}_{OLS}$ is now outperformed by all candidate DR estimators in terms of both bias and MSE.

In the analysis of real, as opposed to simulated data, we do not know a priori whether the features of the joint distribution of (Y, T, X) do or do not favor $\hat{\mu}_{OLS}$. Furthermore, with highly variable "inverse probability" weights, we generally cannot learn the answer from the data, owing to poor power. This suggests that, with highly variable weights, a single estimator, whether $\hat{\mu}_{OLS}$ or a single DR estimator of the mean μ is never adequate even with MAR data; rather an analyst should either "not attempt to make inference about the mean" or else provide a sensitivity analysis (in which models for both the propensity score and the regression of Y on X and estimators of μ are varied). In Section 6, we sketch a possible approach to sensitivity analysis.

In this discussion we ask the following question: can we find DR estimators that, under the authors' chosen joint distribution for (Y, T, X), both perform almost as well as $\hat{\mu}_{OLS}$ applied to data (TY, T, X) and yet perform better than $\hat{\mu}_{OLS}$ when applied to data $((1 - T)Y, T, X)$. In Section 4 we describe the principles we used to search among the set of possible DR estimators and discuss the expected performance of various candidates. We define a general class of DR estimators, which we refer to as "bounded," that contains the DR estimators that perform best in the setting of highly variable "inverse probability" weights. We further subdivide the class of bounded DR estimators into two subclasses—bounded Horvitz–Thompson DR estimators and regression DR estimators. We then describe various scenarios which favor one subclass over the other. We also explain why certain DR estimators perform particularly poorly in settings with highly variable "inverse probability" weights. The performance of our estimators is examined in the simulations reported in Section 5, which both mimics the simulations in S&K and also repeats it but now using data $((1 - T)Y, T, X)$.

A major point emphasized by K&S was that, in their simulations, the regression estimator $\hat{\mu}_{OLS}$ outperformed any DR estimator when both their model for the propensity score and for the regression of Y on X (from now on referred to as the "outcome model") were

misspecified. However, they restricted attention to linear logistic propensity score models. In Section 3, we show that $\hat{\mu}_{OLS}$ is CAN for μ when either (but not necessarily both) a linear model for the inverse propensity score, $1/\pi(X) = X^T\alpha$, or a linear model $X^T\beta$ for the conditional mean $E(Y|X)$ is correctly specified. That is, by definition, $\hat{\mu}_{OLS}$ itself is a DR estimator of μ when the inverse linear model $\pi(X) = 1/(X^T\alpha)$ for the propensity score is substituted for K&S's linear logistic model!

In K&S's simulation experiment, the linear model $X^T\beta$ and the model $\pi(X) = 1/(X^T\alpha)$ are both misspecified. Yet, under their scenario, $\hat{\mu}_{OLS}$ did not "outperform any DR estimator that is CAN when either the regression model $X^T\beta$ or the model $\pi(x) = 1/(x^T\alpha)$ is correctly specified," precisely because $\hat{\mu}_{OLS}$ is one such DR estimator! Of course, the model $\pi(X) = 1/(X^T\alpha)$ would rarely, if ever, be used in practice as the model does not naturally constrain $\pi(X)$ to lie in $[0, 1]$. Nonetheless, understanding that $\hat{\mu}_{OLS}$ is a DR estimator provides important insight into the meaning and theory of double-robustness.

2. GENERAL FORM OF DOUBLE-ROBUST ESTIMATORS

The authors note that many different DR estimators exist and give a number of explicit examples. The authors restrict attention to MAR data with missing response. In this setting Robins (2000), Robins (2002), Tan (2006) and van der Laan and Robins (2003) had previously proposed a rather wide variety of DR estimators, in addition to the DR estimators of Robins and colleagues considered by the authors. Moreover, Scharfstein et al. (1999) and van der Laan and Robins (2003) provided general methods for the construction of DR estimators in models with MAR or coarsened at random (CAR) data. Robins and Rotnitzky (2001) described a general approach to the construction of DR estimators (when they exist) in a very large model class that includes all MAR and CAR models as well as certain nonignorable (i.e., non-CAR) missing data models. Recently, van der Laan and Rubin (2006) have developed a general approach called "targeted maximum likelihood" that has overlap with methods in Scharfstein et al. (1999), Robins (2000, 2002) and Bang and Robins (2005) in the setting of missing response data. We will use the general methods of Robins and Rotnitzky (2001) to find a candidate set of DR estimators among which we then search for ones that perform in simulations as well as or better than those discussed by S&K.

Most of the DR estimators of μ discussed by K&S are of the general form

$$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}) = \mathbb{P}_n\{\hat{m}(X)\} + \mathbb{P}_n\left[\frac{T}{\hat{\pi}(X)}\{Y - \hat{m}(X)\}\right]$$

or

$$\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}) = \mathbb{P}_n\{\hat{m}(X)\} + \frac{\mathbb{P}_n[T/\hat{\pi}(X)\{Y - \hat{m}(X)\}]}{\mathbb{P}_n\{T/(\hat{\pi}(X))\}}, \tag{1}$$

where throughout, $\mathbb{P}_n(A)$ is a shortcut for $n^{-1} \cdot \sum_{i=1}^n A_i$. Robins, Sued, Lei-Gomez and Rotnitzky (2007) show that these estimators are solutions to particular augmented inverse probability weighted (AIPW) estimating equations. The AIPW estimating equations are obtained by applying the general methods of Robins and Rotnitzky (2001) to the simple missing-data model considered by K&S.

Quite generally, to construct $\hat{m}$ and $\hat{\pi}$ we specify (i) a "working" parametric submodel for the propensity score $\pi(X) \equiv \Pr(T = 1|X)$ of the form

$$\pi(\cdot) \in \{\pi(\cdot; \alpha) : \alpha \in \mathbb{R}^q\}, \tag{2}$$

where $\pi(x; \alpha)$ is a known function, for example, $\pi(x; \alpha) = \{1 + \exp(-x^T\alpha)\}^{-1}$ as in S&K and, (ii) a working parametric model for $m(X) \equiv E(Y|X)$ of the form

$$m(\cdot) \in \{m(\cdot; \beta) : \beta \in \mathbb{R}^p\}, \tag{3}$$

where $m(x; \beta)$ is a known function, for example, $m(x; \beta) = x^T\beta$ as in S&K. We then obtain estimators $\hat{\alpha}$ and $\hat{\beta}$ which converge at rate $n^{1/2}$ to some constant vectors α^* and β^*, which are, respectively, equal to the true value of α and/or β when the corresponding working model is correctly specified, and define $\hat{m}(x) \equiv m(x; \hat{\beta})$ and $\hat{\pi}(x) \equiv \pi(x; \hat{\alpha})$. [In fact, under mild additional regularity conditions, the rate $n^{1/2}$ can be relaxed to n^{ζ}, $\zeta > 1/4$, a fact which is critical when the dimensions p and q of β and α are allowed to increase with n as n^{ρ}, $\rho < 1/2$.]

Under regularity conditions, if either (but not necessarily both) (2) or (3) is correct, $\hat{\mu}_{DR}(\hat{\pi}, \hat{m})$ and $\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m})$ are consistent and asymptotically normal (CAN) estimators of μ.

In the special case in which (a) $\pi(x; \alpha) = \{1 + \exp(-x^T\alpha)\}^{-1}$ and $m(x; \beta) = \Phi(x^T\beta)$ where Φ is a known canonical inverse link function, and (b) $\hat{\alpha}$ is the MLE of α and $\hat{\beta}$ is the iteratively reweighted least squares estimator of β among respondents (throughout denoted as $\hat{\beta}_{REG}$) satisfying

$$\mathbb{P}_n[TX\{Y - \Phi(X^T\hat{\beta}_{REG})\}] = 0, \tag{4}$$

we shall denote $\hat{m}$ with $\hat{m}_{REG}$ and the resulting DR estimators as $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$ and $\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$. When Φ is the identity, $\hat{\beta}_{REG}$ is thus the OLS estimator of β. In such case, $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$ is the estimator denoted $\hat{\mu}_{BC\text{-}OLS}$ in S&K and $\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$ is the estimator in the display following (8) in S&K.

3. $\hat{\mu}_{OLS}$ AS A DR ESTIMATOR UNDER A LINEAR INVERSE PROPENSITY MODEL

As anticipated in Section 1, in this section we will argue that the regression estimator $\hat{\mu}_{OLS}$ is indeed a DR estimator with respect to specific working models. Suppose that we postulate the linear inverse propensity model, that is, in (2) we take $\pi(x;\alpha) = 1/(\alpha^T x)$. It follows from (4) that for any $\alpha \in \Omega$,

$$\mathbb{P}_n\left[\frac{T}{\pi(X;\alpha)}\{Y - \hat{m}_{REG}(X)\}\right] = 0. \tag{5}$$

Thus, the regression estimator $\hat{\mu}_{OLS}$ is indeed equal to $\hat{\mu}_{DR}(\pi(\cdot;\alpha), \hat{m}_{REG})$ for any α and therefore DR with respect to the linear inverse propensity model $\mathcal{P}_{\text{sub,inv}}$ and the outcome model $\Phi(x^T\beta) = x^T\beta$.

To estimate α in model $\mathcal{P}_{\text{sub,inv}}$ we may use either the estimator $\hat{\alpha}_{\text{inv}}$ or the estimator $\hat{\hat{\alpha}}_{\text{inv}}$ that, respectively, minimize the log-likelihood $\mathbb{P}_n[T\log\{\pi(X;\alpha)\} + \{1-T\}\log\{1-\pi(X;\alpha)\}]$ or squared norm $\|\mathbb{P}_n[\{\frac{T}{\pi(X;\alpha)} - 1\}X]\|^2$ both subject to the constraints $\pi(X_i;\alpha) \geq 0$, $i = 1, \ldots, N$. Under regularity conditions, $\hat{\alpha}_{\text{inv}}$ and $\hat{\hat{\alpha}}_{\text{inv}}$ converge in probability to quantities α^*_{inv} and α^{**}_{inv} with the property that when model $\mathcal{P}_{\text{sub,inv}}$ is correctly specified, $\pi(X;\alpha^*_{\text{inv}})$ and $\pi(X;\alpha^{**}_{\text{inv}})$ are equal to the propensity score $P(T=1|X)$.

4. DOUBLE-ROBUST ESTIMATORS WITH DESIRABLE PROPERTIES

4.1 Boundedness

We would like to have DR estimators of $\mu = E(Y)$ with the "boundedness" property that, when the sample space of Y is finite, they fall in the parameter space for μ with probability 1. Neither $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$ nor $\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$ has this property. We consider two separate ways to guarantee the "boundedness" property.

First, suppose that we found DR estimators that could be written in the IPW form

$$\mathbb{P}_n\{YT/\hat{\pi}(X)\}/\mathbb{P}_n\{T/\hat{\pi}(X)\} \tag{6}$$

for some nonnegative $\hat{\pi}(\cdot)$. Then the property would hold for such estimators. Specifically, the quantity in the last display is a convex combination of the observed Y-values and thus always lies in the interval $[Y_{\min}, Y_{\max}]$ with endpoints the minimum and maximum observed Y-values. But, $[Y_{\min}, Y_{\max}]$ is included in the parameter space for μ because μ is the population mean of Y.

Note that division by $\mathbb{P}_n\{T/\hat{\pi}(X)\}$ is essential to ensure that (6) is in $[Y_{\min}, Y_{\max}]$. In particular, the Horvitz–Thompson estimator $\hat{\mu}_{HT} = \mathbb{P}_n\{YT/\hat{\pi}(X)\}$ does not satisfy this property. For example, if Y is Bernoulli, (6) lies in $[0,1]$ but $\hat{\mu}_{HT}$ may lie outside $[0,1]$. For instance, this will be the case if in the sample there exists a unit with $T=1$, $\hat{\pi}(X) < 1/n$ and $Y=1$ since then $\hat{\mu}_{HT}$ will be greater than 1. Indeed, when we carried out 1000 Monte Carlo replications of Kang and Schaeffer's simulation experiment for sample size $n = 1000$ using their misspecified analysis model for $\pi(x)$, we found in one particularly anomalous replication, a simulated unit with $T=1$ but with the unusually small estimated propensity $\hat{\pi}(X) < 1/17{,}000$. Thus, had we simulated Y from a Bernoulli rather than from a normal distribution, we would have had $\hat{\mu}_{HT} > 17$ for this anomalous replication!

The desire that an estimator falls in the interval $[Y_{\min}, Y_{\max}]$ is in conflict with the desire that it be unbiased, as we now show. Suppose the propensity score function $\pi(x)$ were known. The set of exactly unbiased estimators of μ (that are invariant to permutations of the index i labeling the units) are contained in the set $\{\mathbb{P}_n[T\{Y - q(X)\}/\pi(X) + q(X)]\}$ as $q(X)$ varies. It follows that no unbiased estimator of μ exists for Y Bernoulli that is guaranteed to fall in $[Y_{\min}, Y_{\max}]$. Taking $q(X)$ identically zero, we obtain $\tilde{\mu}_{HT} = \mathbb{P}_n\{YT/\pi(X)\}$. Suppose that in an actual study of 1000 subjects, a rare fluctuation had occurred and there was a subject with $T=1$ whose propensity score $\pi(X)$ was less than $1/17{,}000$ so $\tilde{\mu}_{HT} > 17$. We doubt any scientist could be convinced to publish the logically impossible estimate $\tilde{\mu}_{HT}$ for the mean of a Bernoulli Y, with the argument that only then would his estimator of the mean be exactly unbiased over hypothetical repetitions of the study that, of course, neither have occurred nor will occur. Exactly analogous difficulties arise for any other choice of $q(X)$. With highly variable weights, "boundedness" trumps unbiasedness.

Second, the boundedness property also holds for DR estimators that can be written in the regression form

$$\mathbb{P}_n\{\hat{m}(X)\} \tag{7}$$

with $\hat{m}(x) = \Phi(x^T\hat{\beta} + h(x)^T\hat{\gamma})$ for some specified function $h(\cdot)$ and an inverse link function Φ satisfying

$\inf \mathbf{Y} \le \Phi(u) \le \sup \mathbf{Y}$ for all u, where $\mathbf{Y}$ is the sample space of Y. This follows because (i) $\mathbb{P}_n\{\hat{m}(X)\}$ falls in the interval $[\hat{m}_{\min}, \hat{m}_{\max}]$, with $\hat{m}_{\min}$ and $\hat{m}_{\max}$ the minimum and maximum values of $\hat{m}(X)$ among the n sample units and, (ii) the above choice of $\Phi(\cdot)$ guarantees that $[\hat{m}_{\min}, \hat{m}_{\max}]$ is contained in the parameter space for μ.

Neither the estimator $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$ nor the estimator $\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$ of Section 2 satisfies the "boundedness" property. Note, however, that $\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m})$ satisfies $|\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m})| < \max_{i=1,\ldots,n} |Y_i - \hat{m}(X_i)| + \max_{i=1,\ldots,n} |\hat{m}(X_i)|$. Thus when Y is Bernoulli and $m(X) = \Phi(X^T\beta)$ for Φ any inverse link with range in $[0, 1]$, we have that $|\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m})| < 2$, so it is within a factor of 2 of lying in the parameter space. In contrast, when Y is Bernoulli, $\hat{\mu}_{DR}(\hat{\pi}, \hat{m})$, like $\hat{\mu}_{HT}$, can be extremely large.

In the next sections, we describe general approaches to constructing DR estimators that can be written in the form (6) or the form (7). Thus it is important to determine whether DR estimators that satisfy (6) perform better or worse than those satisfying (7) when models for both $m(\cdot)$ and $\pi(\cdot)$ are wrong. Unfortunately, no general recommendation can be given because the answer will depend on the specific data generating process and models used to estimate $m(\cdot)$ and $\pi(\cdot)$. For example, suppose that as in Kang and Schaeffer's simulation experiment, Y is continuous, $\text{var}(Y|X) = \sigma^2$ does not depend on X, Φ is the identity link and we estimate the model $X^T\beta$ for $m(X) = E(Y|X)$ by OLS. When (i) $\sigma^2/\text{Var}(Y)$ is near zero and (ii) there exist a number of nonrespondent units j (i.e., units j with $T_j = 0$) whose values x_j of X lie far outside the convex hull of the set of values of X for the subsample of respondents, then y_j and $m(x_j)$ will be close to one another but not to $\hat{m}_{REG}(x_j)$ except if, by luck, the model $E(Y|X) = X^T\beta$ is so close to being correct that the fit of the model to the subsample of respondents allows successful linear extrapolation to X's far from those fitted. As we shall see, it is precisely such "luck" that explains the good performance of $\hat{\mu}_{OLS}$ in the authors' simulation experiment. Without such luck, the estimator $\mathbb{P}_n\{\hat{m}_{REG}(X)\}$ may perform poorly compared to $\mathbb{P}_n\{YT/\hat{\pi}(X)\}/\mathbb{P}_n\{T/\hat{\pi}(X)\}$, owing to unsuccessful linear extrapolation. On the other hand, when $\sigma^2/\text{Var}(Y)$ is close to 1 and very few units with $T = 0$ have values of X far outside the convex hull of the set of X's in the respondents subsample, $\mathbb{P}_n\{\hat{m}_{REG}(X)\}$ will generally perform better than $\mathbb{P}_n\{YT/\hat{\pi}(X)\}/\mathbb{P}_n\{T/\hat{\pi}(X)\}$, as the latter may be dominated by the large weights $1/\hat{\pi}(X)$ assigned to respondents who have both very small values of $\hat{\pi}(X)$ and large residuals $Y - m(X)$.

4.1.1 *Regression double-robust estimators.* We refer to DR estimators satisfying (7) as regression DR estimators. They are obtained by replacing $\hat{m}_{REG}$ in $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$ with $\hat{m}(X)$ satisfying

$$\mathbb{P}_n\left[\frac{T}{\hat{\pi}(X)}\{Y - \hat{m}(X)\}\right] = 0. \tag{8}$$

Here we describe three such estimators, though others exist (Robins et al., 2007).

The first one, proposed in Scharfstein et al. (1999) and discussed further in Bang and Robins (2005), is the estimator in K&S's Table 8. To compute this estimator one considers an extended outcome model of the form $\Phi(X^T\beta + \varphi\hat{\pi}(X)^{-1})$ adding the covariate $\hat{\pi}(X)^{-1}$ (i.e., the inverse of the fitted propensity score). One then jointly estimates (β, φ) with $(\widetilde{\beta}, \widetilde{\varphi})$ satisfying $\mathbb{P}_n[T\{Y - \Phi(X^T\widetilde{\beta} + \widetilde{\varphi}\hat{\pi}(X)^{-1})\}\begin{bmatrix}\hat{\pi}(X)^{-1}\\ X\end{bmatrix}] = 0$. The first row of this last equation is precisely (8) with $\hat{m}(X)$ replaced with the fitted regression function $\hat{m}_{EXT\text{-}REG}(X) = \Phi(X^T\widetilde{\beta} + \widetilde{\varphi}\hat{\pi}(X)^{-1})$. Consequently, $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{EXT\text{-}REG}) = \mathbb{P}_n\{\hat{m}_{EXT\text{-}REG}(X)\}$. This estimator is CAN provided either the model $\pi(x;\alpha)$ for the propensity score $\pi(x)$ or the model $\Phi(x^T\beta)$ for $E(Y|X = x)$ is correct. In fact, it is CAN even if model $\Phi(x^T\beta)$ is incorrect provided the model $\Phi(X^T\beta + \varphi\{\pi(X;\alpha^*)\}^{-1})$ is correct, where α^* is the probability limit of the estimator $\hat{\alpha}$ of α. In particular, as indicated in the previous subsection, if Y is Bernoulli and Φ is the inverse logit link, $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{EXT\text{-}REG})$ is always in $[0, 1]$.

Nonetheless, when Y is continuous and Φ is the identity, $|\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{EXT\text{-}REG})|$ can be disastrously large when the estimated inverse probability weights $\hat{\pi}(X)^{-1}$ are highly variable. Specifically, when $\hat{\pi}(X)^{-1}$ is highly variable, it could very well happen that in most repeated samples the largest value of $\hat{\pi}^{-1}$ among nonrespondents is manyfold greater than the largest value among the respondent subsample. [E.g., a typical Monte Carlo replication of Kang and Schaeffer under the wrong propensity score model had a largest $\hat{\pi}(X)^{-1}$ of 80 in the respondent subsample but a largest $\hat{\pi}(X)^{-1}$ of 1800 in the nonrespondent subsample.] In such cases, enormously greater extrapolation would be required with model $\Phi\{X^T\beta + \varphi\hat{\pi}(X)^{-1}\}$ than with model $\Phi(X^T\beta)$ to obtain fitted values for Y in the nonrespondent subsample, clearly a problem if the extrapolation model $\Phi\{X^T\beta + \varphi\hat{\pi}(X)^{-1}\}$ is also wrong. This phenomenon explains the disastrous performance of

$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{EXT\text{-}REG})$ observed in K&S's Table 8 when both the model for the propensity score and the outcome model are wrong.

A second DR estimator with the regression form (7) is immediately obtained by estimating the parameter β of the model $E(Y|X) = \Phi(X^T\beta)$ with the weighted least squares estimator $\hat{\beta}_{WLS}$ that uses weights $1/\hat{\pi}(X)$. By definition, the estimator $\hat{\beta}_{WLS}$ satisfies

$$\mathbb{P}_n\left[\frac{T}{\hat{\pi}(X)}\{Y - \Phi(X^T\hat{\beta}_{WLS})\}X\right] = 0$$

and consequently (8) is immediately true for $\hat{m}(X)$ equal to $\hat{m}_{WLS}(X) = \Phi(X^T\hat{\beta}_{WLS})$ when, as we always assume in this discussion, the first component of X is the constant 1. It therefore follows that when model $\Phi(X^T\beta)$ has an intercept, $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS}) = \mathbb{P}_n(\hat{m}_{WLS})$. The estimator $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$ is called $\hat{\mu}_{WLS}$ in K&S.

With highly variable $\hat{\pi}(X)^{-1}$ and incorrect models for both $\pi(X)$ and $E(Y|X)$, we would expect $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$ to outperform $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{EXT\text{-}REG})$ because it does not have the severe extrapolation problem of $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{EXT\text{-}REG})$. This expectation is dramatically borne out in K&S's simulations.

Some years ago, Marshall Joffe pointed out to us that $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$ was double-robust and asked us if it had advantages compared to $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{EXT\text{-}REG})$. At the time we had not realized that $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{EXT\text{-}REG})$ would perform so very poorly in settings with highly variable $\hat{\pi}(X)^{-1}$, so we told him that it probably offered no particular advantage. Based on our bad advice, Joffe never published a paper on $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$ as a DR estimator. To our knowledge, Kang and Schaeffer are the first to do so. We note that Kang and Schaeffer do not consider $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$ to be an AIPW DR estimator. However, the above derivation shows otherwise.

Even $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$ may not perform well in some instances. For example, if $\mathrm{Var}(Y|X) = \sigma^2$ is constant, $\sigma^2/\mathrm{Var}(Y)$ is near 1 and a number of nonrespondents have X lying far outside the convex hull of the respondents' X values, then $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$ may perform poorly. This is because the subjects who have the greatest $\hat{\pi}(X)^{-1}$ in the respondents' subsample will have enormous leverage which can force their residual to be nearly zero, which is a problem particularly when $\sigma^2/\mathrm{Var}(Y)$ is near 1 and the model $\Phi(X^T\beta)$ is misspecified, as then extrapolation to the X's far from the convex hull will be poor.

The third DR regression type estimator is an extension of the estimator $\hat{\mu}_{IPW\text{-}NR}$ in Kang and Schaeffer. To compute this estimator we extend the regression model $\Phi(X^T\beta)$ by adding the covariate $\hat{\pi}(X)$ (rather than its inverse) to obtain $\Phi\{X^T\beta + \varphi\hat{\pi}(X)\}$ and then jointly estimate (β, φ) with the estimator $(\widetilde{\beta}, \widetilde{\varphi})$ satisfying

$$\mathbb{P}_n\left[\frac{T}{\hat{\pi}(X)}\{Y - \Phi(X^T\widetilde{\beta} + \widetilde{\varphi}\hat{\pi}(X))\}\begin{bmatrix}\hat{\pi}(X)\\ X\end{bmatrix}\right] = 0.$$

Because we have assumed the vector X has one component equal to the constant 1, (8) is satisfied with $\hat{m}(X)$ equal to $\hat{m}_{DR\text{-}IPW\text{-}NR}(X) = \Phi\{X^T\widetilde{\beta} + \widetilde{\varphi}\hat{\pi}(X)\}$. Thus, $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{DR\text{-}IPW\text{-}NR}) = \mathbb{P}_n(\hat{m}_{DR\text{-}IPW\text{-}NR})$. Furthermore, since by construction, $\mathbb{P}_n[T\{Y - \hat{m}_{DR\text{-}IPW\text{-}NR}(X)\}] = 0$, then $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{DR\text{-}IPW\text{-}NR})$ is also equal to $\mathbb{P}_n\{TY + (1-T)\hat{m}_{DR\text{-}IPW\text{-}NR}(X)\}$. Because $\hat{\pi}(X)$ is bounded between 0 and 1, adding the covariate $\hat{\pi}(X)$ to model $\Phi(X^T\beta)$, in contrast to adding $\hat{\pi}(X)^{-1}$, does not induce model extrapolation problems like the ones discussed above for $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{EXT\text{-}REG})$. We speculate that $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{DR\text{-}IPW\text{-}NR})$ will behave much better than $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{EXT\text{-}REG})$ and possibly similarly to $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$ when $\hat{\pi}(X)^{-1}$ has high variance. Indeed, we have observed this behavior in the simulation study of Section 5; however, due to space limitations, results for $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{EXT\text{-}REG})$ were not reported as they were qualitatively similar to those reported in K&S.

Finally, the estimator $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{EXT\text{-}REG})$ with Φ the identity link is also an example of a DR targeted maximum likelihood estimator of the marginal mean μ in the sense of van der Laan and Rubin (2006). We thus conclude from the above discussion that with highly variable $\hat{\pi}(X)^{-1}$ and incorrect parametric models for $\pi(X)$ and $E(Y|X)$, certain targeted maximum likelihood estimators can perform much worse than the ad hoc estimator $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$.

4.1.2 *Bounded Horvitz–Thompson double-robust estimators.* We refer to DR estimators satisfying (6) as bounded Horvitz–Thompson DR estimators. They are obtained by replacing $\hat{\pi}$ in $\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$ with $\hat{\pi}_{EXT}$ satisfying

$$\mathbb{P}_n\left[(\hat{m}_{REG}(X) - \hat{\mu}_{REG})\left(\frac{T}{\hat{\pi}_{EXT}(X)} - 1\right)\right] = 0 \tag{9}$$

where $\hat{\mu}_{REG} = \mathbb{P}_n[\hat{m}_{REG}(X)]$. We can obtain such a $\hat{\pi}_{EXT}(\cdot)$ by considering the extended logistic model $\pi_{EXT}(X) = \mathrm{expit}\{\alpha^T X + \varphi h(X)\}$ with $h(X)$ a user-supplied function and estimating φ with $\widetilde{\varphi}_{PROP\text{-}GREED}$ solving

$$\mathbb{P}_n\left[\left\{\frac{T}{\mathrm{expit}(\hat{\alpha}^T X + \varphi h(X))} - 1\right\} \cdot \{\hat{m}_{REG}(X) - \hat{\mu}_{OLS}\}\right] = 0$$

where $\hat{\alpha}$ is the MLE of α in the model $\pi(X) = \text{expit}(\alpha^T X)$. Then $\hat{\pi}_{EXT}(X) = \text{expit}\{\hat{\alpha}^T X + \tilde{\varphi}_{PROP\text{-}GREED} h(X)\}$ satisfies (9) and consequently $\hat{\mu}_{B\text{-}DR}(\hat{\pi}_{EXT}, \hat{m}_{REG})$ is of the form (6). A default choice for $h(X)$ would be $\hat{m}_{REG}(X) - \hat{\mu}_{OLS}$.

Interestingly, the OLS estimator $\hat{\mu}_{OLS}$ can be viewed not only as a DR estimator, as seen in Section 3, but also as a bounded Horvitz–Thompson DR estimator! Specifically, suppose that $\hat{\hat{\alpha}}_{\text{inv}}$is used to estimate α in the propensity model $\pi(X;\alpha)$ as in Section 3, except that without imposing the constraints, so that indeed, $\hat{\hat{\alpha}}_{\text{inv}}$ solves $0 = \mathbb{P}_n[\{\frac{T}{\pi(X;\alpha)} - 1\}X]$. Then

$$\mathbb{P}_n\left[\left\{\frac{T}{\pi(X;\hat{\hat{\alpha}}_{\text{inv}})} - 1\right\}(\hat{m}_{REG}(X) - \hat{\mu}_{OLS})\right] = 0$$

and $\hat{\mu}_{OLS}$ is equal to the inverse probability weighted estimator

$$\mathbb{P}_n\left\{\frac{T}{\pi(X;\hat{\hat{\alpha}}_{\text{inv}})}Y\right\}\Big/\mathbb{P}_n\left\{\frac{T}{\pi(X;\hat{\hat{\alpha}}_{\text{inv}})}\right\}.$$

Robins (2001) and van der Laan and Rubin (2006) describe particular bounded Horvitz–Thompson DR estimators $\hat{\mu}^{(\infty)}$ that are obtained by iterating to convergence a sequence $\hat{\mu}^{(j)}$ of estimators that are not themselves bounded Horvitz–Thompson estimators. However, the Robins (2001) estimator performed poorly in our simulations (results not shown) and no simulation study of the van der Laan and Rubin (2006) estimator has been published to our knowledge with highly variable inverse probability weights. In fact, van der Laan and Rubin (2006) describe an estimator $\hat{\mu}^{(\infty)}$ that is simultaneously a bounded Horvitz–Thompson and a regression DR estimator that is obtained by iterating to convergence a sequence $\hat{\mu}^{(j)}$ of estimators without this dual property. Again we do not know of a simulation study showing that the estimator generally performs well in practice with highly variable inverse probability weights.

5. SIMULATION STUDIES

To investigate the nature of the surprisingly good performance of the regression estimator $\hat{\mu}_{OLS}$ in the simulation study of K&S and to evaluate the performance of the additional estimators described in Section 4, we replicated the simulation study of K&S. Table 1 reports the Monte Carlo bias, variance and mean squared error for twelve different estimators of $\mu = 210$, sample sizes $n = 200$ and 1000 and the four possible combinations of model specifications for the propensity score and the conditional mean of the response given the covariates (the latter referred to throughout as the outcome model).

The estimators reported in rows 1 and 9, rows 3 and 11, rows 4, 12, 17 and 22 and rows 5, 13, 18 and 23 are, respectively, the estimators $\hat{\mu}_{OLS}$, $\hat{\mu}_{IPW\text{-}POP}$, $\hat{\mu}_{BC\text{-}OLS}$ and $\hat{\mu}_{WLS}$ investigated by K&S. Throughout we use the notational conventions of Sections 2–4, and thus we rename $\hat{\mu}_{BC\text{-}OLS}$ and $\hat{\mu}_{WLS}$ with $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$ and $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$, respectively. The estimator $\hat{\mu}_{HT}$ is the Horvitz–Thompson type estimator $\mathbb{P}_n\{TY/\hat{\pi}(X)\}$. All remaining estimators are DR estimators of μ and are defined in Sections 2 and 4.

When both working models are correct, theory indicates that all DR estimators are CAN, asymptotically equivalent, and efficient in the class of estimators that are CAN even if the outcome model is incorrect. We were not surprised then to find that all DR estimators reported in rows 4–8 of Table 1 performed identically and were more efficient than the inefficient IPW estimators $\hat{\mu}_{IPW\text{-}POP}$ and $\hat{\mu}_{HT}$. However, the near-identical behavior of the regression estimator $\hat{\mu}_{OLS}$ caught our attention. The estimator $\hat{\mu}_{OLS}$ is the maximum likelihood estimator of μ, and hence efficient, in a semiparametric model that assumes a parametric form for the conditional mean of Y given the covariates. Thus, we would have expected it to have smaller variance than that of the DR estimators of μ, because when both the propensity score model and the regression model are correct, the latter attains the semiparametric variance bound in the less restrictive (nonparametric) model that does not impose restrictions on the conditional mean of Y. A closer examination of the data generating process used by K&S explains this unusual behavior. Under S&K data generating process $Y_i = 210 + 13.7Z_i^* + \varepsilon_i$, where $Z_i^* = 2Z_{1i} + \sum_{j=2}^4 Z_{ji}$, with Z_{ji}, $j = 1, \ldots, 4$, and ε_i mutually independent $N(0, 1)$ random variables. But under this process Z_i^*, and hence $Z_i = (Z_{1i}, Z_{2i}, Z_{3i}, Z_{4i})$, is an essentially perfect predictor of Y_i: the residual variance $\text{var}(Y_i|Z_i^*)$ is equal to $\text{var}(Y_i)/195$. This striking feature of K&S data generating process is illustrated in Figure 1. The figure shows a scatterplot of Y versus the predicted values from the fit of the correct outcome model to the respondents in a random sample of $n = 200$ units. Dark dots correspond to data points of respondents. White dots correspond to the simulated missing outcomes Y_i of the nonrespondents plotted against the predicted values $Z_i'\hat{\beta}$. The white dots follow nearly perfectly a straight line through the origin and

TABLE 1
Results for simulation study as in K&S

Row	Estimator	Sample size 200			Sample size 1000		
		Bias	Var	MSE	Bias	Var	MSE
Both models right							
1	$\hat{\mu}_{OLS}$	0.13	5.97	5.98	−0.03	1.41	1.41
2	$\hat{\mu}_{HT}$	−0.08	148.92	148.92	0.17	26.46	26.49
3	$\hat{\mu}_{IPW\text{-}POP}$	−0.06	14.12	14.13	−0.03	3.43	3.43
4	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$	0.13	5.96	5.98	−0.03	1.41	1.41
5	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$	0.13	5.97	5.98	−0.03	1.41	1.41
6	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{DR\text{-}IPW\text{-}NR})$	0.13	5.97	5.98	−0.03	1.41	1.41
7	$\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$	0.13	5.96	5.98	−0.03	1.41	1.41
8	$\hat{\mu}_{B\text{-}DR}(\hat{\pi}_{EXT}, \hat{m}_{REG})$	0.13	5.97	5.98	−0.03	1.41	1.41
Both models wrong							
9	$\hat{\mu}_{OLS}$	−0.39	10.91	11.06	−0.83	2.19	2.88
10	$\hat{\mu}_{HT}$	16.87	4110.86	4395.39	38.97	39933	41452
11	$\hat{\mu}_{IPW\text{-}POP}$	1.67	73.39	76.17	4.81	108.86	131.95
12	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$	−4.90	145.93	169.91	−13.91	6853.68	7047.12
13	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$	−2.01	10.70	14.74	−2.98	2.20	11.08
14	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{DR\text{-}IPW\text{-}NR})$	−1.76	11.82	14.90	−2.49	1.81	8.02
15	$\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$	−3.82	40.07	54.65	−8.03	128.61	193.13
16	$\hat{\mu}_{B\text{-}DR}(\hat{\pi}_{EXT}, \hat{m}_{REG})$	−2.25	11.77	16.82	−3.33	3.44	14.54
π-model right, outcome model wrong							
17	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$	0.55	11.82	12.12	0.07	2.81	2.82
18	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$	0.65	8.82	9.24	0.16	1.90	1.93
19	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{DR\text{-}IPW\text{-}NR})$	0.06	7.39	7.40	−0.10	1.58	1.59
20	$\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$	0.56	11.51	11.83	0.08	2.79	2.80
21	$\hat{\mu}_{B\text{-}DR}(\hat{\pi}_{EXT}, \hat{m}_{REG})$	0.53	9.41	9.69	0.11	2.08	2.09
π-model wrong, outcome model right							
22	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$	0.14	5.95	5.97	−0.03	1.77	1.77
23	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$	0.13	5.97	5.98	−0.03	1.41	1.41
24	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{DR\text{-}IPW\text{-}NR})$	0.13	5.97	5.98	−0.03	1.41	1.41
25	$\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$	0.13	5.96	5.97	−0.02	1.43	1.43
26	$\hat{\mu}_{B\text{-}DR}(\hat{\pi}_{EXT}, \hat{m}_{REG})$	0.13	5.96	5.98	−0.02	1.42	1.42

with slope 1: the predicted values are essentially perfect predictors of the missing outcomes! When the outcome and propensity score models are correctly specified, the asymptotic variance of the DR estimator is equal to $\text{var}(Y) + \text{var}[\pi(Z)\{1 - \pi(Z)\}^{-1}\,\text{var}(Y|Z)]$. When Z is a perfect predictor of Y, this variance reduces to $\text{var}(Y)$, the variance of the standardized distribution of $\hat{\mu}_{FULL}$, the sample mean of Y of respondents and nonrespondents. This is not surprising because it is well known that, when the outcome model is correctly specified, a DR estimator asymptotically extracts all the information available in Z to predict Y. Since the regression estimator $\hat{\mu}_{OLS}$ cannot be more efficient than $\hat{\mu}_{FULL}$, we conclude that $\hat{\mu}_{OLS}$ and the DR estimators should have nearly identical variance when Z is an almost perfect predictor of Y and indeed this variance should be also almost the same as that of the infeasible estimator $\hat{\mu}_{FULL}$. In our study we had simulated the outcomes of the nonrespondents. Thus, we were indeed able to compute $\hat{\mu}_{FULL}$ and its Monte Carlo variance. As expected, the Monte Carlo variance of $\hat{\mu}_{FULL}$ was essentially the same as that of $\hat{\mu}_{OLS}$ for both sample sizes.

Theory also indicates that the IPW estimators $\hat{\mu}_{IPW\text{-}POP}$ and $\hat{\mu}_{HT}$ of rows 2 and 3 should be CAN. However, in our simulations, these estimators were nearly unbiased but their sampling distribution was skewed to the right and had very large variance. Figure 2 shows smooth density estimators for these sampling distributions for sample sizes $n = 200$ and $n = 1000$. The skewness and large variance of the IPW estimators were caused by few samples which had respondents with large values of Y and very large weights $1/\hat{\pi}$. Specifically, in most samples, the true π values of

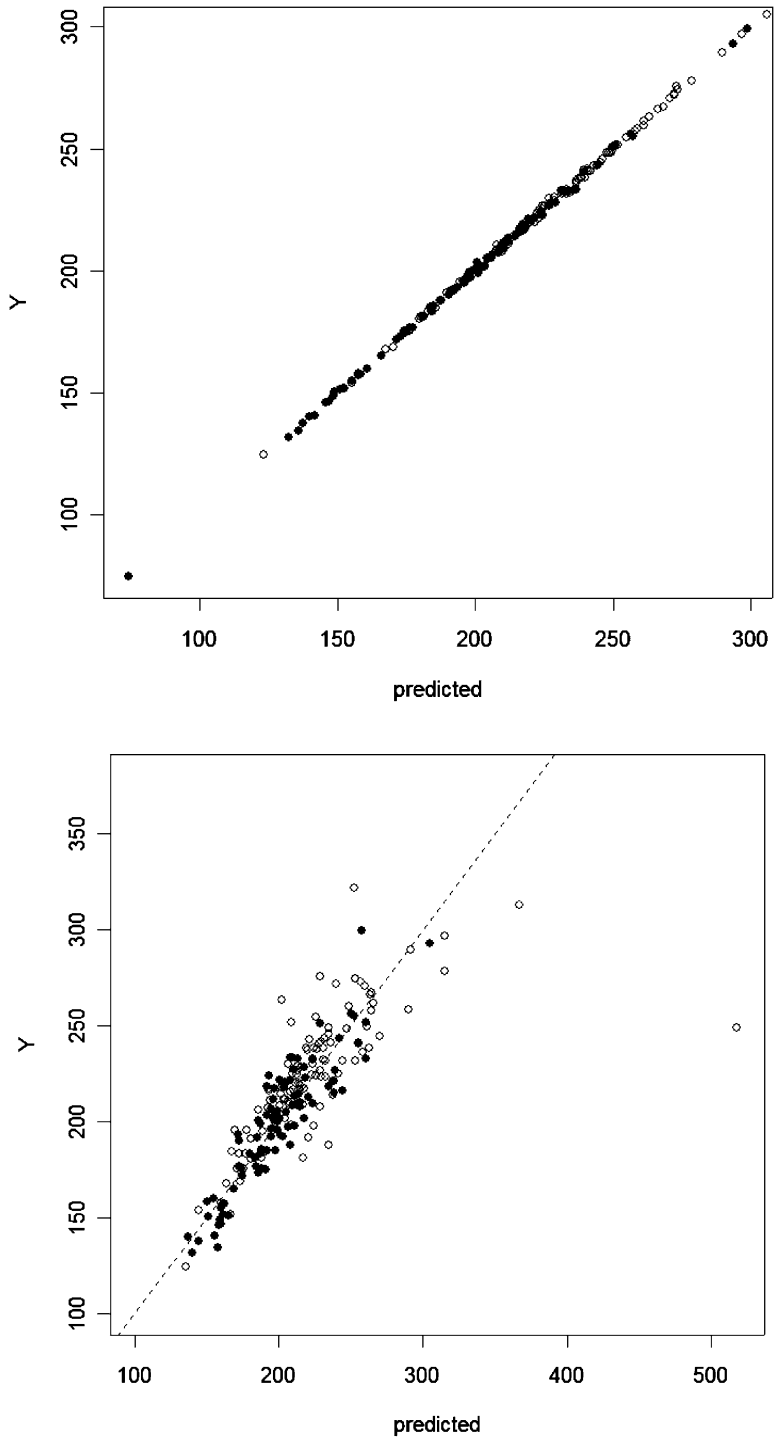


FIG. 1. *K&S simulation experiment. Outcomes vs predicted values. Sample size* 200. *Top*: *correct y model. Bottom*: *wrong y model. Dashed line is the line* $Y = X$. *Dark dots*: *respondents. White dots*: *nonrespondents.*

the respondents were not too small, and consequently the weights $1/\hat{\pi}$ not too large, precisely because by the very definition of π, having a respondent with a small π is a rare event. In the data generating process of K&S, $\pi(Z)$ is negatively correlated with Y; the correlation is roughly equal to -0.6. Thus, in most samples, the $1/\hat{\pi}$-weighted mean of the Y values of the respondents tended to be smaller than μ. However, in a few samples, some anomalous respondent had a small value of π. In the computation of $\hat{\mu}_{HT}$, this anomalous respondent carried an unusually large weight $1/\hat{\pi}$ and because his Y value tended to be larger than the mean μ, the estimator $\hat{\mu}_{HT}$ in those rare samples tended to be substantially larger than μ. The skewness lessens as the sample size increases because with large sample sizes, the number of samples which have respondents with small values of π also increases. The skewness is also substantially less severe for $\hat{\mu}_{IPW\text{-}POP}$ compared to that of $\hat{\mu}_{HT}$ also as expected since, as discussed in Section 4.1, in any given sample, $|\hat{\mu}_{IPW\text{-}POP}|$ is bounded by the largest observed $|Y|$ value.

Although the Monte Carlo sampling distribution of the IPW estimators gives a rough idea of the shape of the true sampling distribution of these estimators, neither the Monte Carlo bias nor the Monte Carlo variance should be trusted. One thousand replications are not enough to capture the tail behavior of highly skewed sampling distributions, and as such cannot produce reliable Monte Carlo estimates of bias, much less of variance.

Turn now to the case in which the propensity score model is correct but the outcome model is incorrect. Theory indicates that the DR estimators of rows 17 to 21 of Table 1 should be CAN. However, in our simulations nearly all the DR estimators were slightly biased upward. Nevertheless, all DR estimators performed as well as or better, in terms of MSE, than the OLS estimator of row 9.

Consider now the case in which the propensity score model is incorrect but the outcome model is correct. Once again, the almost identical performance of all DR estimators in rows 22–26 of Table 1 with that of the OLS estimator of row 1 is no surprise after recalling that Z is a perfect predictor of Y. Specifically, the fact that Z_i^* is a nearly perfect predictor of Y_i implies that $\hat{m}(Z_i)$ is almost identical to the outcome Y_i regardless of whether unit i is a respondent or a nonrespondent and regardless of whether $\hat{m}(Z_i)$ was fit by ordinary least squares or by weighted least squares. Thus, the average of $\hat{m}(Z_i)$ is essentially the same as $\hat{\mu}_{FULL}$ and the sum of $T_i\hat{\pi}_i^{-1}(Y_i - \hat{m}(Z_i))$ is almost zero regardless of the model under which $\hat{\pi}$ was computed. Consequently, all DR estimators must be nearly the same as the infeasible full data sample mean $\hat{\mu}_{FULL}$.

Finally, turn to the case in which both propensity score and the outcome models are wrong. The performance of the IPW estimators is disastrous as well as that of the DR estimator in row 12 and, to a lesser extent, that of the estimator in row 15. Figure 3 shows smooth density estimators of the sampling distribution of these four estimators when the sample size is 1000. The estimators $\hat{\mu}_{HT}$ and $\hat{\mu}_{IPW\text{-}POP}$ have distributions heavily skewed to the right while the estimators $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$ and $\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$ have distributions heavily skewed to the left. The skewness is far more dramatic for the estimators $\hat{\mu}_{HT}$ and $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$ than for their counterparts $\hat{\mu}_{IPW\text{-}POP}$ and $\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$, reflecting the fact that $\hat{\mu}_{IPW\text{-}POP}$ and $\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$ are bounded in the sense described in Section 4.1 while $\hat{\mu}_{HT}$ and $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$ are unbounded. [Indeed, to avoid distortions, in constructing the density plots of $\hat{\mu}_{HT}$ and $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$ we have omitted the extreme values of 5873 and -2213, respectively, from one simulation replication.] Rows 12 and 14 of Table 1 report that the Monte Carlo bias and variance indeed are even larger for $n = 1000$ than for $n = 200$. The extreme distribution skewness and the increase in bias and variance with sample size are explained as follows. As noted earlier, even when the π's are estimated from a correct model, the distribution of $\hat{\mu}_{HT}$ and $\hat{\mu}_{IPW\text{-}POP}$ will tend to be skewed to the right when $1/\pi$ is positively correlated with Y because of the presence of a few unusual samples with anomalous respondents with large Y values and small π values. Now, because of the nature of the wrong analytic propensity score model used in the simulations, the estimated $\hat{\pi}$'s corresponding to the anomalous units in the unusual samples were many times smaller than the true π's. As a consequence the, usually large, values of Y of the anomalous units essentially determined the values of $\hat{\mu}_{HT}$ and $\hat{\mu}_{IPW\text{-}POP}$ in the unusual samples and consequently, exacerbated even more the skewness of the Monte Carlo sampling distribution of the IPW estimators. The larger bias and variance when $n = 1000$ than when $n = 200$ were due to two replications with sample size 1000 in which the values of the estimators were extreme [specifically, $\hat{\mu}_{HT} = 1475$ and 2884, and $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG}) = -2213$ and -175]. These outlying values were caused by one anomalous nonrespondent in each sample with large values of Y (the second largest Y values in one sample and the

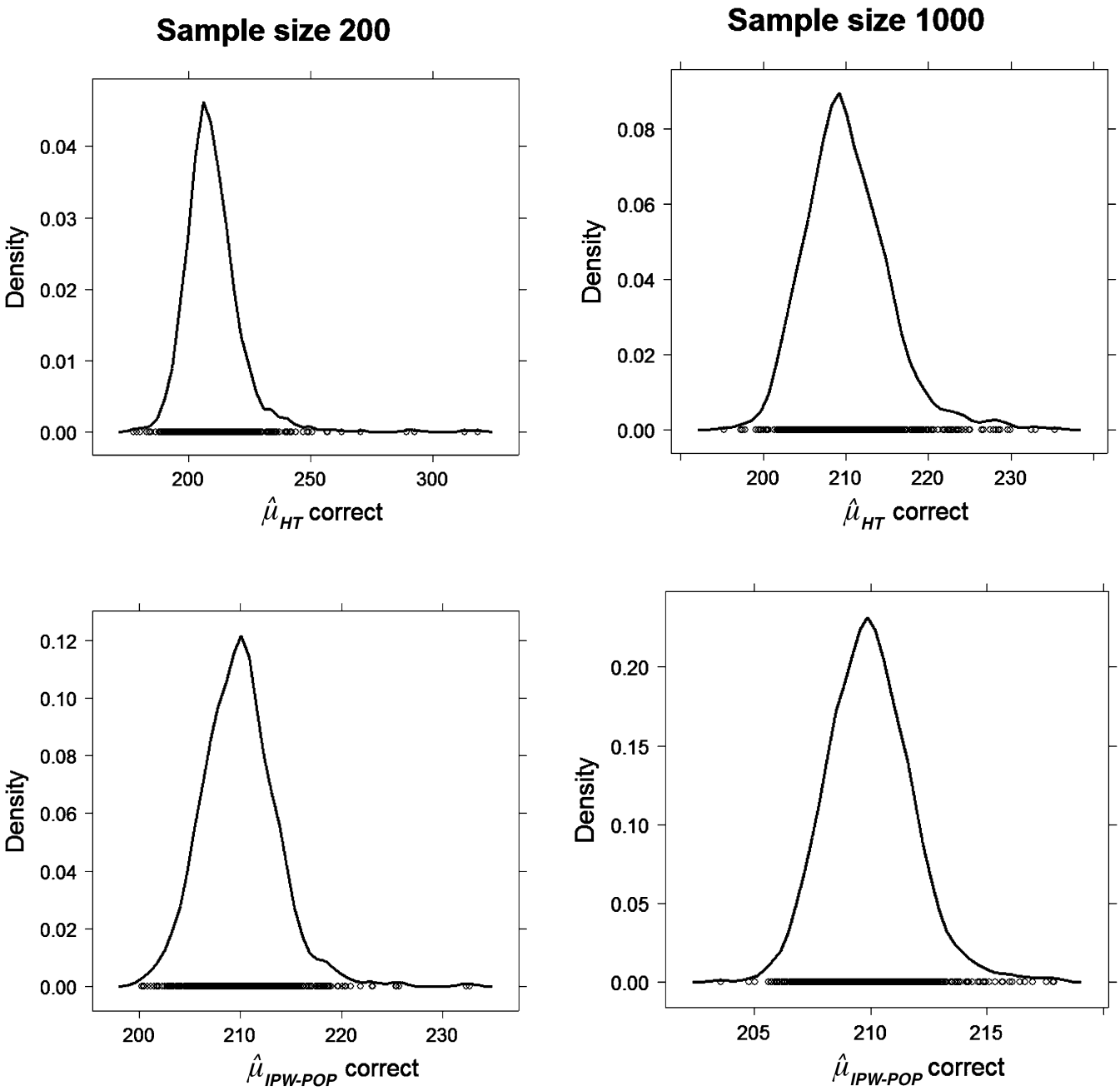


FIG. 2. *Distributions of $\hat{\mu}_{HT}$ and $\hat{\mu}_{IPW\text{-}POP}$ under correct propensity score models.*

largest in the other). For these units, the $1/\pi$ values were 38.7 and 50.9 but $1/\hat{\pi}$ were 17,068 and 399, respectively. When the two samples with these anomalous units were removed, the variance of the estimators $\hat{\mu}_{HT}$ and $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$ decreased to 6729 and 890, respectively. The paradoxical increase in Monte Carlo variance with sample size is but another proof that the Monte Carlo variance in simulations with 1000 replications is not a reliable estimator of the true variance for estimators with highly skewed distributions. The different directionality of the skewness of the IPW and DR estimators is explained as follows. In the computation of $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$ and $\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$ we inverse probability weight the values of $(\hat{m}_{REG} - Y)$. Consequently since, as indicated below, under K&S's wrong analytic outcome model, $\hat{m}_{REG}$ was reasonably bounded; thus, in the few unusual samples, the anomalous units with small π's had large and negative values of $(\hat{m}_{REG} - Y)$ and produced extremely small values of the DR estimators.

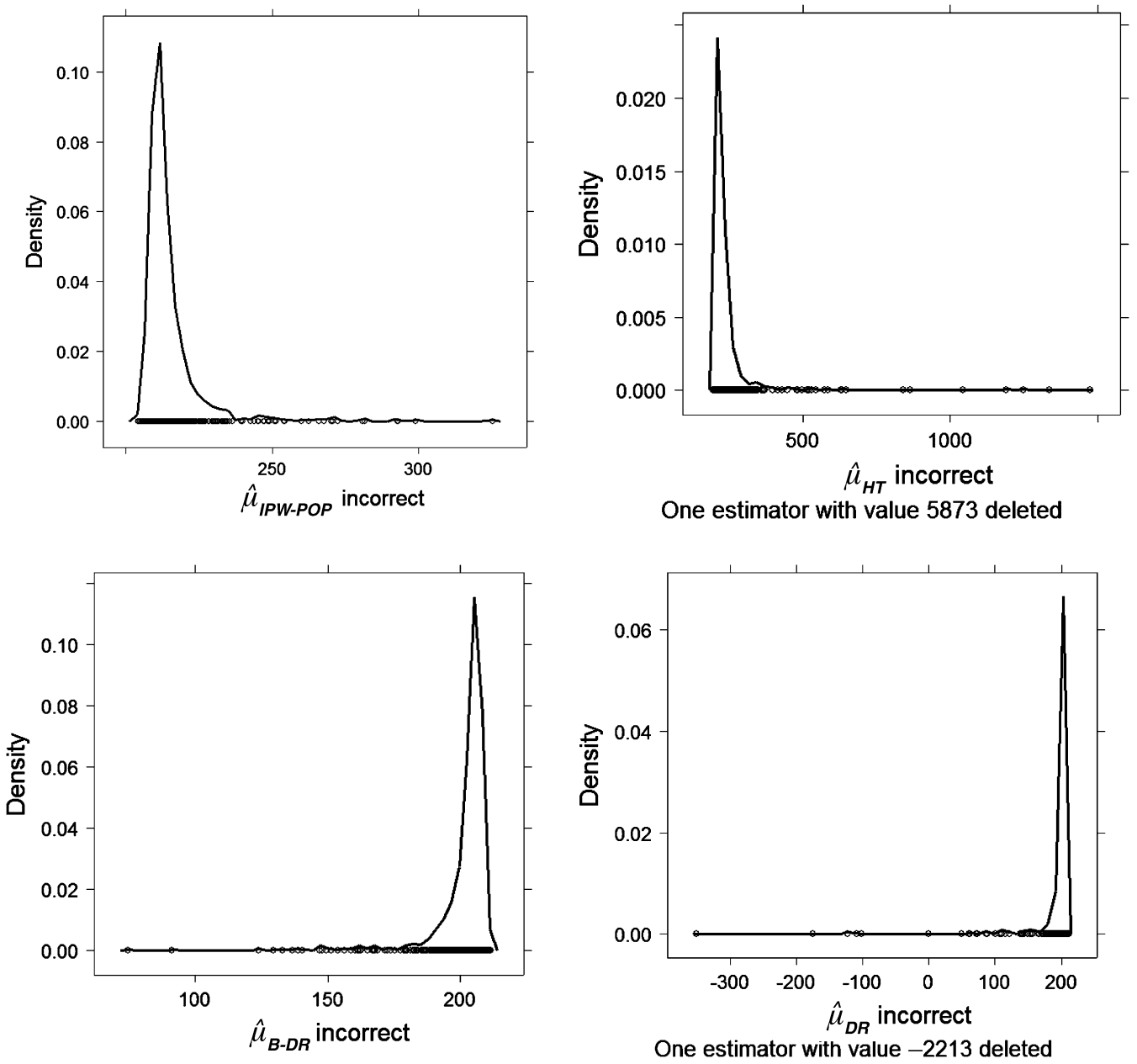


FIG. 3. *Distributions of* $\hat{\mu}_{HT}, \hat{\mu}_{IPW\text{-}POP}, \hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$ *and* $\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$ *under incorrect propensity score and outcome models.*

The performance of the remaining DR estimators in rows 13, 14 and 16 is heterogeneous. Some, though still biased, have bias and variance orders of magnitude smaller than the variance of the estimators $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$ and $\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$.

In a second simulation experiment described below, the relative performance of the DR estimators was somewhat different than in this simulation study and, as we explain later, better than that of the regression estimator $\hat{\mu}_{OLS}$. This attests to the obvious fact that when the propensity score and outcome models are both incorrect we cannot expect to find a single clear winner. The relative performance of the estimators will very much depend on the data generating process and the nature of the model misspecifications.

To understand why the regression estimator $\hat{\mu}_{OLS}$ performed so remarkably well when both models were wrong, we first note that because the outcome model was a linear regression model with an intercept fitted by ordinary least squares in the respondent subsample, the sum of the predicted values $X'\hat{\beta}$ and the sum of Y in the respondent subsample are the same. Thus,

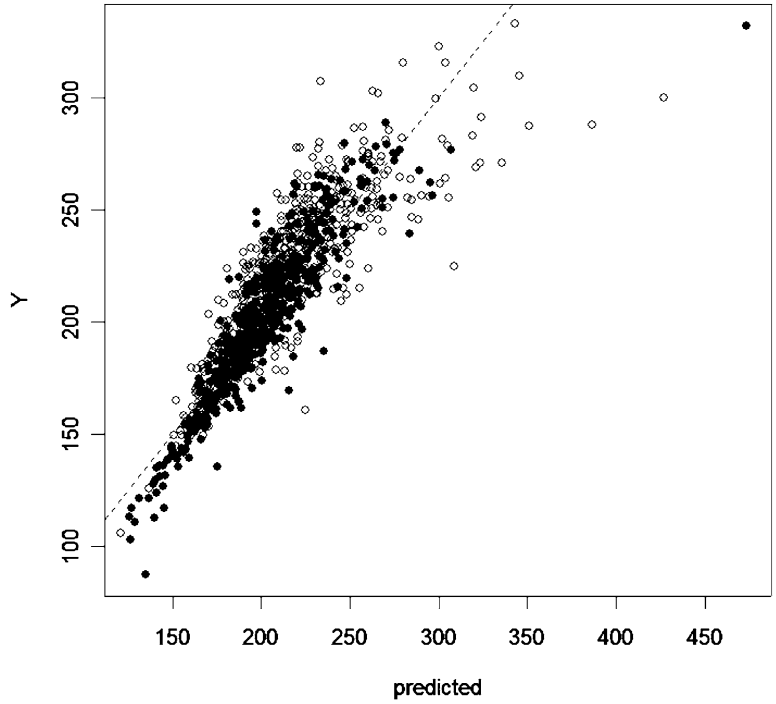


FIG. 4. *Y vs predicted values in one sample of size* 1000 *generated under K&S experiment. Dashed line is the line* $Y = X$. *Dark dots*: *respondents*. *White dots*: *nonrespondents*.

$\hat{\mu}_{OLS} = (n_{\text{obs}}/n)\overline{Y}_{\text{obs}} + (n_{\text{miss}}/n)\ \overline{(X'\hat{\beta})}_{\text{miss}}$, where $\overline{(X'\hat{\beta})}_{\text{miss}}$ is the average of the predicted values for the missing outcomes. The bias of $\hat{\mu}_{OLS}$ therefore depends on the bias of $\overline{(X'\hat{\beta})}_{\text{miss}}$ as an estimator of the mean of Y in the nonrespondent subpopulation. If, due either to good luck or "cherry picking," the prediction function $x'\hat{\beta}$ from a misspecified regression model $x'\beta$ successfully extrapolates to the covariates of nonrespondents, even when these are far from the convex hull of covariates in the respondent subsample, $\overline{(X'\hat{\beta})}_{\text{miss}} - \overline{Y}_{\text{miss}}$ will be roughly centered around 0, and consequently $\hat{\mu}_{OLS}$ will be a nearly unbiased estimator of the mean of Y. We now show this phenomenon explains the excellent performance of $\hat{\mu}_{OLS}$ in Kang & Schaeffer's simulation. In Figure 4 we plotted the outcomes Y versus the predicted values $X'\hat{\beta}$ in the previously mentioned unusual sample of size 1000 with both the propensity and outcome models misspecified, where $\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$, $\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$ and the IPW estimators did disastrously due to the presence of one anomalous unit with extremely small $\hat{\pi}$. The dark dots correspond to the observed data values of the respondents. White dots correspond to the actual simulated missing outcomes Y of the nonrespondents plotted against the predicted values $X'\hat{\beta}$. We can see that the predicted values of the nonrespondents are reasonably centered around the straight line even for those points with predicted values far from the predicted values of the respondents. In this sample, $\hat{\mu}_{OLS}$ was 205.78, a far more reasonable value than those obtained for the IPW and just-mentioned DR estimators.

To demonstrate that $\hat{\mu}_{OLS}$ can have a substantially worse performance than the DR estimators, we conducted a second simulation experiment. This second experiment, like our first, redid K&S's simulation by generating the data (Y, T, X) from K&S's chosen distributions. However, in our second experiment we analyzed the data $((1-T)Y, T, X)$ in which Y is observed only when $T = 0$, rather than the data (TY, T, X) that was analyzed by us in our first experiment and by K&S in their paper. [To do so, since the data $((1-T)Y, T, X)$ can be recoded as $((1-T)Y, 1-T, X)$, we simply recompute each of the estimators reported in Table 1 except now we everywhere replace $\hat{\pi}$ and T by $1-\hat{\pi}$ and $1-T$.] The results are displayed in Table 2. We observe that, with both models wrong, the bias and MSE of $\hat{\mu}_{OLS}$ now exceed those of any DR estimator!

As in our first experiment, due to the extreme variability in the estimated "inverse probability" weights, the DR estimators appear to have considerable finite

TABLE 2
Results for simulation study as in K&S but with the roles of T and $1 - T$ *reversed*

Row	Estimator	Sample size 200			Sample size 1000		
		Bias	Var	MSE	Bias	Var	MSE
Both models right							
1	$\hat{\mu}_{OLS}$	0.12	5.96	5.98	−0.03	1.41	1.41
2	$\hat{\mu}_{HT}$	−0.46	49.14	49.36	−0.24	8.45	8.51
3	$\hat{\mu}_{IPW\text{-}POP}$	0.45	14.76	14.96	0.05	3.11	3.12
4	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$	0.12	5.96	5.98	−0.02	1.41	1.41
5	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$	0.12	5.96	5.97	−0.02	1.41	1.41
6	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{DR\text{-}IPW\text{-}NR})$	0.12	5.96	5.97	−0.02	1.40	1.41
7	$\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$	0.12	5.96	5.97	−0.02	1.41	1.41
8	$\hat{\mu}_{B\text{-}DR}(\hat{\pi}_{EXT}, \hat{m}_{REG})$	0.12	5.96	5.97	−0.02	1.40	1.41
Both models wrong							
9	$\hat{\mu}_{OLS}$	4.97	7.97	32.68	4.97	1.91	26.62
10	$\hat{\mu}_{HT}$	0.55	40.27	40.57	0.39	6.27	6.43
11	$\hat{\mu}_{IPW\text{-}POP}$	3.92	9.67	25.03	3.68	2.22	15.79
12	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$	3.33	8.79	19.90	3.07	2.12	11.53
13	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$	3.17	8.21	18.24	2.81	1.97	9.84
14	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{DR\text{-}IPW\text{-}NR})$	3.11	8.21	17.90	2.64	1.97	8.94
15	$\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$	3.32	8.70	19.69	3.04	2.10	11.34
16	$\hat{\mu}_{B\text{-}DR}(\hat{\pi}_{EXT}, \hat{m}_{REG})$	3.30	8.68	19.55	3.01	2.10	11.16
π-model right, outcome model wrong							
17	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$	0.71	12.60	13.11	0.14	2.96	2.98
18	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$	0.99	8.04	9.02	0.23	1.92	1.97
19	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{DR\text{-}IPW\text{-}NR})$	0.71	7.26	7.76	0.18	1.72	1.75
20	$\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$	0.75	11.21	11.76	0.14	2.76	2.78
21	$\hat{\mu}_{B\text{-}DR}(\hat{\pi}_{EXT}, \hat{m}_{REG})$	0.86	10.38	11.12	0.18	2.71	2.74
π-model wrong, outcome model right							
22	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{REG})$	0.12	5.96	5.97	−0.02	1.40	1.41
23	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{WLS})$	0.12	5.96	5.97	−0.02	1.40	1.41
24	$\hat{\mu}_{DR}(\hat{\pi}, \hat{m}_{DR\text{-}IPW\text{-}NR})$	0.12	5.96	5.97	−0.02	1.40	1.41
25	$\hat{\mu}_{B\text{-}DR}(\hat{\pi}, \hat{m}_{REG})$	0.12	5.96	5.97	−0.02	1.40	1.41
26	$\hat{\mu}_{B\text{-}DR}(\hat{\pi}_{EXT}, \hat{m}_{REG})$	0.12	5.96	5.97	−0.02	1.40	1.41

sample bias, especially at the smaller sample size of 200, when the propensity model is correct but the outcome model is wrong. In fact, this bias is larger than it was in the first simulation experiment, which was to be expected as the variability in the estimated "inverse probability" weights was greater in the second than the first experiment (data not shown).

6. SENSITIVITY ANALYSIS

Consider again the missing-data setting with the mean μ of Y as the parameter of interest. When the covariate vector X is high dimensional, one cannot be certain, owing to lack of power, that a chosen model for the propensity score is nearly correct, even if it passes standard goodness-of-fit tests. Therefore a large number of models for the propensity score with different subsets of the covariates, different orders of interactions and different dimensions of the parameter vector should be fit to the data. Similarly, many different outcome models should be fit. This raises the question: once fit, how should these many candidate models be used in the estimation of the mean of Y?

One approach is to use modern techniques of model selection to choose a single propensity and outcome model. Specifically, there has been a recent outpouring of work on model selection in regression. This work has shown that one can use cross-validation and/or penalization to empirically choose, from among a large number of candidates, a model whose predictive risk for the response variable in the regression is close to that of the best candidate model. In fact, van der Laan (2005) has proposed that k-fold cross-validation should be routinely employed to select the model for

the propensity score and for the outcome regression that are to be used in the construction of a DR estimator.

An alternative approach which we are currently studying is the following. Suppose one has fit J_p propensity score models and J_o outcome models. For a favorite DR estimator $\hat{\mu}$, define $\hat{\mu}_{ij}$ as the DR estimator that uses the fitted values from the ith propensity model and the jth outcome model. Now, if the ith propensity model is correct, all J_o estimators in the set $\mathcal{E}_{p,i} \equiv \{\hat{\mu}_{ij}; j = 1, \ldots, J_o\}$ will be CAN estimators of μ. Thus, an α-level test of the homogeneity hypothesis $H_{pi}: E^A(\hat{\mu}_{i1}) = E^A(\hat{\mu}_{ij})$ for all $j \in \{2, \ldots, J_o\}$ [where $E^A(\cdot)$ stands for large sample mean, i.e., the probability limit of $\cdot$] is also an α-level goodness-of-fit test for the propensity model that is directly relevant to its use in a DR estimator of μ. Similarly if the jth outcome model is correct, all J_p estimators in the set $\mathcal{E}_{o,j} \equiv \{\hat{\mu}_{ij}; i = 1, \ldots, J_p\}$ will be CAN for μ and a test of the homogeneity hypothesis $H_{oj}: E^A(\hat{\mu}_{1j}) = E^A(\hat{\mu}_{ij})$ for all $i \in \{2, \ldots, J_p\}$ is a test of fit for the outcome model. This suggests that one could choose as a final estimator of μ the DR estimator $\hat{\mu}_{i^*j^*}$, where i^* is the i for which the test of the hypothesis H_{pi} gave the largest p-value and j^* is the j for which the test of the hypothesis H_{oj} gave the largest p-value. However, this method of selecting i^* and j^* is nonoptimal for two reasons. First, it could easily select a misspecified propensity model i for which the power of the test of the hypothesis H_{pi} is particularly poor and similarly for the outcome regression. This remark implies that some measure of the spread of the elements of $\mathcal{E}_{p,i}$ and $\mathcal{E}_{o,j}$ should also contribute to the selection of i^* and j^*. Second, the method does not exploit the fact that if i^* and j^* are correct, then $E^A(\hat{\mu}_{ij^*}) = E^A(\hat{\mu}_{i^*j})$ for all i and j, suggesting that an optimal method should select i^* and j^* jointly. Alternative approaches for selecting i^* and j^* will be reported elsewhere. In any case, the very fact that input to the selection algorithm requires the matrix $\hat{\mu}_{ij}$ provides an informal sensitivity analysis; we directly observe the sensitivity of our DR estimator to the choice of propensity and outcome regression model.

The approach just described could also be combined with the model selection approach. Specifically, one first uses cross-validation to choose not one but rather J_p and J_o propensity and outcome models (the ones with the J_p and J_o lowest cross-validated risk estimates) out of a much larger number of candidate models and next, one uses these $J_p + J_o$ models as input for the approach described above. Sensitivity to the choice of the particular DR estimator might be included by using a number of different DR estimators and selecting among or averaging over DR estimators that give similar estimates $\hat{\mu}_{i^*j^*}$.

van der Laan (2005) has proposed some new approaches to model selection for DR estimation that go beyond his above-mentioned approach, which we do not discuss here due to space limitations. Finally, Wang, Petersen, Bangsberg and van der Laan (2006) have proposed using the parametric bootstrap to study the sensitivity of DR estimates to highly variable "inverse probability" weights.

7. FURTHER CONSIDERATIONS

Estimation of Causal Effects

K&S briefly touch on the problem of estimating the difference of the outcome means corresponding to two treatments in an observational study under ignorability. This difference is often referred to as the average causal effect (ACE). K&S view the problem of estimating ACE essentially as two missing-data problems, each one regarding the outcomes of subjects that do not follow the treatment of concern as missing. The difference of the DR estimators of the separate means serves as an estimator of the mean difference ACE. However, the difference of the two DR estimators will have poor small sample behavior if there is incomplete overlap of the estimated propensity scores in the treated and untreated. In fact, in the presence of incomplete overlap, most investigators argue against trying to estimate ACE and in favor of estimating the causal effect in the subpopulation of subjects with overlapping propensity scores. However, assuming the ACE parameter is of some substantive interest, Robins et al. (2007) suggest an alternative to reporting the difference of two DR estimators of the separate means. Their approach is based on fitting a linear semiparametric regression model for the unknown conditional effect function encoding the dependence of the conditional treatment effect on the baseline covariates X. Their model has the property that it is guaranteed to be correctly specified under the null hypothesis that the conditional effect function is the zero function. Robins et al. (2007) show that this strategy results in estimators of the ACE that greatly outperform any estimator based on the difference of double-robust estimators, whenever the model for the conditional effect function is correctly specified; in particular, when the aforementioned null hypothesis is true.

Multiple Robustness

Consider again the MAR missing-data model with X very high dimensional (say 20–100 continuous covariates) so we cannot possibly hope to model the propensity score or the outcome regression nonparametrically. Double-robust estimators of the mean μ of Y are $n^{1/2}$-consistent if either one of two parametric models is correct but inconsistent if both models are misspecified. This property of DR estimators seems unsatisfactory, as it means that one does very, very well if one of the two models is correct but can do very, very poorly when both are incorrect. Might we do better?

Define an estimator to be *m-robust* for μ at rate n^{α} if the estimator is n^{α}-consistent for μ when any one of m parametric models is correct, but inconsistent if all m models are misspecified. A DR estimator is then a 2-robust estimator with $\alpha = 1/2$. Our view is that an m-robust estimator with m large, even though this may require α to be much smaller than $1/2$ and so entail a much slower rate of convergence, would usually be preferable to a DR estimator for the following two reasons. First, if one uses an m-robust rather than a DR estimator, one is more likely to be using a consistent estimator of μ (as it is always more likely that at least one of m, rather than one of two, models is correct). Second, the slower rate of convergence (under the assumption one of the m models is correct) will result in wider nominal confidence intervals than the usual nominal intervals of length $1/n^{1/2}$ associated with a DR estimator. Such a wide interval seems to us a more appropriate measure of the actual uncertainty about μ, more accurately reflecting the fact that our estimator could even be inconsistent if all m models are incorrect.

These observations raise the following questions. Do *m-robust* estimators exist for arbitrarily large m if we are willing to sacrifice $n^{1/2}$*-consistency* for n^{α}-consistency with α perhaps much smaller than $1/2$? What is the maximum value of α we can achieve for a given m? If *m-robust* estimators exist for $m > 2$, how do we construct them? Answers to these questions can be found in Robins, Li, Tchetgen and van der Vaart (2007), where it is shown that, under weak assumptions, (i) *m-robust* estimators exist for all m, (ii) m-robust estimators are $(m-1)$ dimensional U-statistics, for which explicit closed-form expressions are given, and (iii) the maximal possible α is often less than $1/2$ and sometimes much less. However, the finite sample properties of *m-robust* estimators have yet to be studied even by simulation. Thus we will have to wait to see if they are as useful in practice as theory would indicate they should be.

REFERENCES

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189

ROBINS, J. M. (2000). Robust estimation in .sequentially ignorable missing data and casual inference models. *Proc. Amer. Statist. Assoc. Section on Bayesian Statistical Science* **1999** 6–10.

ROBINS, J. M. (2002). Commentary on "Using inverse weighting and predictive inference to estimate the effects of time-varying treatments on the discrete-time hazard", by Dawson. and Lavori. *Statistics in Medicine* **21** 1663–1680.

ROBINS, J. M. and ROTNITZKY, A. (2001). Comment on "Inference for semiparametric models: Some questions and an answer," by P. J. Bickel and J. Kwon. *Statist. Sinica* **11** 920–936. MR1867326

ROBINS, J. M. and WANG, N. (2000). Inference for imputation estimators. *Biometrika* **87** 113–124. MR1766832

ROBINS, J. M., ROTNITZKY, A. and ZHAO L.-P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90** 106–121. MR1325118

ROBINS, J. M., LI, L., TCHETGEN, F. and VAN DER VAART, A. W. (2007). Higher order influence functions and minimax estimation of nonlinear functionals. In *IMS Lecture Notes–Monograph Series Probability and Statistics Models*: *Essays in Honor of David A. Freedman* **2** 335–421.

ROBINS, J. M., SUED, M., LEI-GOMEZ, Q. and ROTNITZKY, A. (2007). Double-robustness with improved efficiency in missing and causal inference models. Technical report, Harvard School of Public Health.

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. Springer, New York. MR1899138

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric non-response models (with discussion). *J. Amer. Statist. Assoc.* **94** 1096–1146. MR1731478

TAN, Z. (2006). A distributional approach for causal inference using propensity scores. *J. Amer. Statist. Assoc.* **101** 1619–1637. MR2279484

VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer, New York. MR1958123

VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. U.C. Berkeley Division of Biostatistics Working Paper Series. Available at http://www.bepress.com/ucbbiostat/paper213/.

VAN DER LAAN, M. (2005). Statistical inference for variable importance. U.C. Berkeley Division of Biostatistics Working Paper Series. Available at http://www.bepress.com/ucbbiostat/paper188/.

WANG, Y., PETERSEN, M., BANGSBERG, D. and VAN DER LAAN, M. (2006). Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment. U.C. Berkeley Division of Biostatistics Working Paper Series. Available at http://www.bepress.com/ucbbiostat/paper211/.

Statistical Science
2007, Vol. 22, No. 4, 560–568
DOI: 10.1214/07-STS227A
Main article DOI: 10.1214/07-STS227
© Institute of Mathematical Statistics, 2007

Comment: Understanding OR, PS and DR

Zhiqiang Tan

We congratulate Kang and Schafer (KS) on their excellent article comparing various estimators of a population mean in the presence of missing data, and thank the Editor for organizing the discussion. In this communication, we systematically examine the propensity score (PS) and the outcome regression (OR) approaches and doubly robust (DR) estimation, which are all discussed by KS. The aim is to clarify and better our understanding of the three interrelated subjects.

Sections 1 and 2 contain the following main points, respectively.

(a) OR and PS are two approaches with different characteristics, and one does not necessarily dominate the other. The OR approach suffers the problem of *implicitly* making extrapolation. The PS-weighting approach tends to yield large weights, *explicitly* indicating uncertainty in the estimate.

(b) It seems more constructive to view DR estimation in the PS approach by incorporating an OR model rather than in the OR approach by incorporating a PS model. Tan's (2006) DR estimator can be used to improve upon any initial PS-weighting estimator with both variance and bias reduction.

Finally, Section 3 presents miscellaneous comments.

1. UNDERSTANDING OR AND PS

For a population, let X be a vector of (pretreatment) covariates, T be the treatment status, Y be the observed outcome given by $(1-T)Y_0 + TY_1$, where (Y_0, Y_1) are potential outcomes. The observed data consist of independent and identically distributed copies (X_i, T_i, Y_i), $i = 1, \ldots, n$. Assume that T and (Y_0, Y_1) are conditionally independent given X. The objective is to estimate

$$\mu_1 = E(Y_1),$$
$$\mu_0 = E(Y_0),$$

and their difference, $\mu_1 - \mu_0$, which gives the average causal effect (ACE). KS throughout focused on the problem of estimating μ_1 from the data (X_i, T_i, T_iY_i), $i = 1, \ldots, n$, only, noting in Section 1.2 that estimation of the ACE can be separated into independent estimation of the means μ_1 and μ_0. We shall in Section 3 discuss subtle differences between causal inference and solving two separate missing-data problems, but until then we shall restrict our attention to estimation of μ_1 from (X_i, T_i, T_iY_i) only.

The model described at this stage is completely nonparametric. No parametric modeling assumption is made on either the regression function $m_1(X) = E(Y|T = 1, X)$ or the propensity score $\pi(X) = P(T = 1|X)$. Robins and Rotnitzky (1995) and Hahn (1998) established the following fundamental result for semiparametric (or more precisely, nonparametric) estimation of μ_1.

PROPOSITION 1. *Under certain regularity conditions, there exists a unique influence function, which hence must be the efficient influence function, given by*

$$\begin{aligned}\tau_1 &= \frac{T}{\pi(X)}Y - \mu_1 - \left(\frac{T}{\pi(X)} - 1\right)m_1(X)\\ &= m_1(X) - \mu_1 + \frac{T}{\pi(X)}\big(Y - m_1(X)\big).\end{aligned}$$

The semiparametric variance bound (i.e., the lowest asymptotic variance any regular estimator of μ_1 can achieve) is $n^{-1}E^2(\tau_1)$.

The semiparametric variance bound depends on both $m_1(X)$ and $\pi(X)$. The bound becomes large or even infinite *whenever* $\pi(X) \approx 0$ for some values of X. Intuitively, it becomes difficult to infer the overall mean of Y_1 in this case, because very few values of Y_1 are observed among subjects with $\pi(X) \approx 0$. The difficulty holds *whatever* parametric approach, OR or PS, is taken for inference, although the symptoms can be different. This point is central to our subsequent discussion.

The problem of estimating μ_1 is typically handled by introducing parametric modeling assumptions on either $m_1(X)$ or $\pi(X)$. The OR approach is to specify an

Zhiqiang Tan is Assistant Professor, Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, 615 North Wolfe Street, Baltimore, Maryland 21205, USA (e-mail: ztan@jhsph.edu).

OR model, say $m_1(X;\alpha)$, for $m_1(X)$ and then estimate μ_1 by

$$\hat{\mu}_{OR} = \frac{1}{n}\sum_{i=1}^{n} \hat{m}_1(X_i),$$

where $\hat{m}_1(X)$ is the fitted response. The PS approach is to specify a PS model, say $\pi(X;\gamma)$, for $\pi(X)$ and then estimate μ_1 by

$$\hat{\mu}_{IPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i Y_i}{\hat{\pi}(X_i)}$$

or

$$\sum_{i=1}^{n} \frac{T_i Y_i}{\hat{\pi}(X_i)} \Big/ \sum_{i=1}^{n} \frac{T_i}{\hat{\pi}(X_i)},$$

where $\hat{\pi}(X)$ is the fitted propensity score. The idea of inverse probability weighting (IPW) is to recover the joint distribution of (X, Y_1) by attaching weight $\propto \hat{\pi}^{-1}(X_i)$ to each point in $\{(X_i, Y_i) : T_i = 1\}$ (see Tan, 2006, for a likelihood formulation). More generally, consider the following class of augmented IPW estimators $\hat{\mu}_{AIPW} = \hat{\mu}_{AIPW}(h)$ depending on a *known* function $h(X)$:

$$\hat{\mu}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{1}{n}\sum_{i=1}^{n} \left(\frac{T_i}{\hat{\pi}(X_i)} - 1\right) h(X_i).$$

A theoretical comparison of the two approaches is given by

PROPOSITION 2. *Assume that an OR model is correctly specified and $m_1(X)$ is efficiently estimated with adaptation to heteroscedastic* $\mathrm{var}(Y_1|X)$, *and that a PS model is correctly specified and $\pi(X)$ may or may not be efficiently estimated. Then*

$$\text{asy.var}\,(\hat{\mu}_{OR}) \leq \text{asy.var}\,(\hat{\mu}_{AIPW}),$$

where asy.var. *denotes asymptotic variance as $n \to \infty$.*

In fact, the asymptotic variance of $\hat{\mu}_{OR}$, which is the lowest under the parametric OR model, is no greater than the semiparametric variance bound under the nonparametric model, whereas that of $\hat{\mu}_{AIPW}$ is no smaller than $n^{-1}E^2(\tau_1)$ because τ_1 has the smallest variance among $\pi^{-1}(X)TY - (\pi^{-1}(X)T - 1)h(X)$ over all functions $h(X)$. In the degenerate case where $m_1(X)$ and $\pi(X)$ are known, the comparison can be attributed to Rao–Blackwellization because $E[\pi^{-1}(X)TY - (\pi^{-1}(X)T - 1)h(X)|X] = m_1(X)$. This result has interesting implications for understanding the two approaches.

First, the result formalizes the often-heard statement that the (A)IPW estimator is no more efficient than the OR estimator. If a correct OR model and a correct PS model were placed in two black boxes, respectively, and if a statistician were asked to open one and only one box, then the statistician should choose the box for the OR model in terms of asymptotic efficiency (minus the complication due to adaptation to heteroscedastic variance of Y_1 given X). However, one could immediately argue that this comparison is only of phantom significance, because all models (by human efforts) are wrong (in the presence of high-dimensional X) and therefore the hypothetical situation never occurs. In this sense, we emphasize that the result does *not* establish any absolute superiority of the OR approach over the PS approach.

Second, even though not implying one approach is better than the other, the result does shed light on different characteristics of the two approaches as an approximation to the ideal nonparametric estimation. Typically, increasingly complicated but nested parametric models can be specified in either approach to reduce the dependency on modeling assumptions. For a sequence of OR models, the asymptotic variance of $\hat{\mu}_{OR}$ is increasing to the semiparametric variance bound, whereas for a sequence of PS models, the asymptotic variance of $\hat{\mu}_{AIPW}$ is decreasing to the semiparametric variance bound. For this difference, we suggest that the OR approach is aggressive and the PS approach is conservative. Correctly specifying an OR model ensures that $\hat{\mu}_{OR}$ is consistent and has asymptotic variance no greater, whereas correctly specifying a PS model ensures that $\hat{\mu}_{AIPW}$ is consistent and has asymptotic variance no smaller, than otherwise would be best attained without any modeling assumption. This interpretation agrees with the finding of Tan (2006) that the OR approach works directly with the usual likelihood, whereas the PS approach retains part of all information and therefore ignores other part on the joint distributions of covariates and potential outcomes.

Now the real, hard questions facing a statistician are:

(a) Which task is more likely to be accomplished, to correctly specify an OR model or a PS model?

(b) Which mistake (even a mild one) can lead to worse estimates, misspecification of an OR model or a PS model?

First of all, it seems that no definite comparison is possible, because answers to both questions depend on unmeasurable factors such as the statistician's effort and

experience for question (a) and the degree and direction of model misspecification for question (b). Nevertheless, some informal comparisons are worth considering.

Regarding question (a), a first answer might be "equally likely," because both models involve the same vector of explanatory variables X. However, the two tasks have different forms of difficulties. The OR-model building works on the "truncated" data $\{(X_i, Y_i) : T_i = 1\}$ within treated subjects. Therefore, any OR model relies on extrapolation to predict $m_1(X)$ at values of X that are different from those for most treated subjects [i.e., $\pi(X) \approx 0$]. The usual model checking is not capable of detecting OR-model misspecification, whether mild or gross, in this region of X. (Note that finding high-leverage observations can point to the existence of such a region of X, *not* model misspecification.) This problem holds for low- or high-dimensional X, and is separate from the difficulty to capture $m_1(X)$ within treated subjects when X is high-dimensional [cf. KS's discussion below display (2)]. In contrast, the PS-model building works on the "full" data $\{(X_i, T_i)\}$ and does not suffer the presence of data truncation, although suffering the same curse of dimensionality. The exercise of model checking is capable of detecting PS-model misspecification. The matter of concern is that successful implementation is difficult when X is high-dimensional.

Regarding question (b), KS (Section 2.1) suggested that the (A)IPW estimator is sensitive to misspecification of the PS model when $\pi(X) \approx 0$ for some values of X. For example, if $\pi(X) = 0.01$ is underestimated at 0.001, then, even though the absolute bias is small ($= 0.009$), the weight $\pi^{-1}(X)$ is overestimated by 10 times higher. In this case, the estimator has inflated standard error, which can be much greater than its bias. In contrast, if the OR model is misspecified, then the bias of the OR estimator is the average of those of $\hat{m}_1(X)$ across individual subjects in the original scale, and can be of similar magnitude to its standard deviation.

In summary, OR and PS are two approaches with different characteristics. If an OR model is correctly specified, then the OR estimator is consistent and has asymptotic variance no greater than the semiparametric variance bound. Because of data truncation, any OR model suffers the problem of *implicitly* making extrapolation at values of X with $\pi(X) \approx 0$. Finding high-leverage observations in model checking can point to the existence of such values of X. In contrast, the PS approach specifically examines $\pi(X)$ and addresses data truncation by weighting to recover the joint distribution of (X, Y_1). The weights are necessarily large for treated subjects with $\pi(X) \approx 0$, in which case the standard error is large, *explicitly* indicating uncertainty in the estimate. If a PS model is correctly specified, then the (A)IPW estimator is consistent and has asymptotic variance no smaller than the semiparametric variance bound.

2. UNDERSTANDING DR

The OR or the (A)IPW estimator requires specification of an OR or a PS model, respectively. In contrast, a DR estimator uses the two models in a manner such that it remains consistent if either the OR or the PS model is correctly specified. The prototypical DR estimator of Robins, Rotnitzky and Zhao (1994) is

$$\begin{aligned}\hat{\mu}_{AIPW,\mathrm{fix}} &= \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{\pi}(X_i)} \\ &\quad - \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{T_i}{\hat{\pi}(X_i)} - 1\Big)\hat{m}_1(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\hat{m}_1(X_i) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\hat{\pi}(X_i)}\big(Y - \hat{m}_1(X_i)\big).\end{aligned}$$

The two equivalent expressions [resp. (9) and (8) in KS] correspond to those for the efficient influence function τ_1 in Proposition 1. Proposition 3 collects theoretical comparisons between the three estimators.

PROPOSITION 3. *The following statements hold*:

(i) $\hat{\mu}_{AIPW,\mathrm{fix}}$ *is doubly robust*.

(ii) $\hat{\mu}_{AIPW,\mathrm{fix}}$ *is locally efficient*: *if a PS and an OR model are correctly specified, then* $\hat{\mu}_{AIPW,fix}$ *achieves the semiparametric variance bound and hence*

$$\text{asy.var}\,(\hat{\mu}_{AIPW,\mathrm{fix}}) \;\le\; \text{asy.var}\,(\hat{\mu}_{AIPW}).$$

(iii) *If an OR model is correctly specified and* $m_1(X)$ *is efficiently estimated in* $\hat{\mu}_{OR}$, *then*

$$\text{asy.var}\,(\hat{\mu}_{AIPW,\mathrm{fix}}) \;\ge\; \text{asy.var}\,(\hat{\mu}_{OR}).$$

Compared with $\hat{\mu}_{OR}$, $\hat{\mu}_{AIPW,\mathrm{fix}}$ is more robust in terms of bias if the OR model is misspecified but the PS model is correctly specified, but is less efficient in terms of variance if the OR model is correctly specified. The usual bias-variance trade-off takes effect. Compared with $\hat{\mu}_{AIPW}$, $\hat{\mu}_{AIPW,\mathrm{fix}}$ is more robust in

terms of bias if the PS model is misspecified but the OR model is correctly specified, and is more efficient in terms of variance if both the PS and the OR models are correctly specified. The usual bias-variance trade-off seems not to exist. Intuitively, the difference can be attributed to the characteristics of OR (being aggressive) and PS (being conservative) discussed in Section 1. It is possible for the PS approach to reduce both bias and variance by incorporating an OR model, but not so for the OR approach by incorporating a PS model.

Local efficiency implies that if the PS model is correctly specified, then $\hat{\mu}_{AIPW,\text{fix}}$ gains efficiency over $\hat{\mu}_{AIPW}$ for every function $h(X)$ under the condition that the OR model is also correctly specified. A more desirable situation is to find an estimator that is not only doubly robust and locally efficient but also, whenever the PS model is correctly specified, guaranteed to gain efficiency over $\hat{\mu}_{AIPW}$ for any initial, fixed function $h(X)$. For simplicity, consider $\hat{\mu}_{IPW}$ corresponding to $h(X) = 0$ as the initial estimator. In this case, consider Tan's (2006) regression (tilde) estimator

$$\tilde{\mu}_{REG} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{\pi}(X_i)} - \tilde{\beta}^{(1)}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\hat{\pi}(X_i)} - 1\right)\hat{m}_1(X_i),$$

where $\tilde{\beta}^{(1)}$ is the first element of $\tilde{\beta} = \tilde{E}^{-1}(\hat{\xi}\hat{\zeta}^{\top})\tilde{E}(\hat{\xi}\hat{\eta})$, $\tilde{E}$ denotes sample average, and

$$\hat{\eta} = \frac{TY}{\hat{\pi}(X)},$$

$$\hat{\xi} = \left(\frac{T}{\hat{\pi}(X)} - 1\right)\left(\hat{m}_1(X), \frac{\frac{\partial\hat{\pi}}{\partial\gamma^{\top}}(X)}{1-\hat{\pi}(X)}\right)^{\top},$$

$$\hat{\zeta} = \frac{T}{\hat{\pi}(X)}\left(\hat{m}_1(X), \frac{\frac{\partial\hat{\pi}}{\partial\gamma^{\top}}(X)}{1-\hat{\pi}(X)}\right)^{\top}.$$

This estimator algebraically resembles Robins, Rotnitzky and Zhao's (1995) regression (hat) estimator

$$\hat{\mu}_{REG} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{\pi}(X_i)} - \hat{\beta}^{(1)}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\hat{\pi}(X_i)} - 1\right)\hat{m}_1(X_i),$$

where $\hat{\beta}^{(1)}$ is the first element of $\hat{\beta} = \tilde{E}^{-1}(\hat{\xi}\hat{\xi}^{\top})\tilde{E}(\hat{\xi}\hat{\eta})$. Compared with $\hat{\mu}_{AIPW,\text{fix}}$, each estimator introduces an estimated regression coefficient, $\tilde{\beta}$ or $\hat{\beta}$, of $\hat{\eta}$ against control variates $\hat{\xi}$. Therefore, $\tilde{\mu}_{REG}$ and $\hat{\mu}_{REG}$ share the advantage of optimally using control variates $\hat{\xi}$ [Proposition 4(ii)]. See Section 3 for a discussion about "control variates" and "regression estimators." On the other hand, $\hat{\beta}$ is defined in the classical manner, whereas $\tilde{\beta}$ is specially constructed by exploiting the structure of control variates $\hat{\xi}$. This subtle difference underlies Proposition 4(i).

PROPOSITION 4. *The following statements hold*:

(i) *$\tilde{\mu}_{REG}$ and $\hat{\mu}_{REG}$ are locally efficient, but $\tilde{\mu}_{REG}$ is doubly robust and $\hat{\mu}_{REG}$ is not.*

(ii) *If a PS model is correctly specified and $\pi(X)$ is efficiently estimated, then $\tilde{\mu}_{REG}$ and $\hat{\mu}_{REG}$ achieve the smallest asymptotic variance among*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{\pi}(X_i)} - b^{(1)}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\hat{\pi}(X_i)} - 1\right)\hat{m}_1(X_i),$$

where $b^{(1)}$ is an arbitrary coefficient. The two estimators are asymptotically at least as efficient as $\hat{\mu}_{IPW}$ and $\hat{\mu}_{AIPW,\text{fix}}$, corresponding to $b^{(1)} = 0$ and 1.

Compared with $\hat{\mu}_{AIPW,\text{fix}}$, $\tilde{\mu}_{REG}$ provides a more concrete improvement upon $\hat{\mu}_{IPW}$ due to the possession of three properties: optimality in using control variates, local efficiency and double robustness. Using $\tilde{\mu}_{REG}$ achieves variance reduction if the PS model is correctly specified (the effect of which is maximal if the OR model is also correctly specified), and bias reduction if the PS model is misspecified but the OR model is correctly specified. On the other hand, comparison between $\hat{\mu}_{OR}$ and $\tilde{\mu}_{REG}$ is similarly subject to the usual bias-variance trade-off as that between $\hat{\mu}_{OR}$ and $\tilde{\mu}_{AIPW,\text{fix}}$. That is, $\tilde{\mu}_{REG}$ is more robust than $\hat{\mu}_{OR}$ if the OR model is misspecified but the PS model is correctly specified, but is less efficient if the OR model is correctly specified.

The preceding comparisons between $\hat{\mu}_{AIPW,\text{fix}}$, $\tilde{\mu}_{REG}$ and $\hat{\mu}_{OR}$, $\hat{\mu}_{IPW}$ present useful facts for understanding DR estimation. It seems more meaningful to consider $\hat{\mu}_{AIPW,\text{fix}}$ or $\tilde{\mu}_{REG}$ as an advance or improvement in the PS approach by incorporating an OR model rather than in the OR approach by incorporating a PS model. The OR and PS models play different roles, even though the models are equally referred to in the concept of DR and $\hat{\mu}_{AIPW,\text{fix}}$ can be expressed as bias-corrected $\hat{\mu}_{OR}$ or equivalently as bias-corrected $\hat{\mu}_{IPW}$. This viewpoint is also supported by the construction of $\hat{\mu}_{AIPW,\text{fix}}$ (in the first expression by Robins, Rotnitzky and Zhao, 1994) and $\tilde{\mu}_{REG}$. Both of the estimators are derived under the assumption that the PS model is correct, and then examined in the situation where the OR model is

also correct, or the PS model is misspecified but the OR model correct (see Tan, 2006, Section 3.2).

The different characteristics discussed in Section 1 persist between the PS (even using $\hat{\mu}_{AIPW,\text{fix}}$ or $\tilde{\mu}_{REG}$ with the DR benefit) and OR approaches. The asymptotic variance of $\hat{\mu}_{AIPW}$, $\hat{\mu}_{AIPW,\text{fix}}$, or $\tilde{\mu}_{REG}$ if $\tilde{\mu}_{REG}$ the PS model is correctly specified is no smaller, whereas that of $\hat{\mu}_{OR}$ if the OR model is correctly specified is no greater, than the semiparametric variance bound. Moreover, if the OR model is correct, the asymptotic variance of $\hat{\mu}_{AIPW,\text{fix}}$ or $\tilde{\mu}_{REG}$ is still no smaller than that of $\hat{\mu}_{OR}$. Therefore:

PROPOSITION 5. *The asymptotic variance of $\hat{\mu}_{AIPW,\text{fix}}$ or $\tilde{\mu}_{REG}$ if either a PS or an OR model is correctly specified is no smaller than that of $\hat{\mu}_{OR}$ if the OR model is correctly specified and $m_1(X)$ is efficiently estimated in $\hat{\mu}_{OR}$.*

Like Proposition 2, this result does *not* establish absolute superiority of the OR approach over the PS-DR approach. Instead, it points to considering practical issues of model specification and consequences of model misspecification. There seems to be no definite comparison, because various, unmeasurable factors are involved. Nevertheless, the points regarding questions (a) and (b) in Section 1 remain relevant.

In summary, it seems more constructive to view DR estimation in the PS approach by incorporating an OR model rather than in the OR approach by incorporating a PS model. The estimator $\tilde{\mu}_{REG}$ provides a concrete improvement upon $\hat{\mu}_{IPW}$ with both variance and bias reduction in the sense that it gains efficiency whenever the PS model is correctly specified (and maximally so if the OR model is also correctly specified), and remains consistent if the PS model is misspecified but the OR model is correctly specified. On the other hand, comparison between $\tilde{\mu}_{REG}$ and $\hat{\mu}_{OR}$ is complicated by the usual bias-variance trade-off. Different characteristics are associated with the OR and the PS-DR approaches and should be carefully weighed in applications.

3. OTHER COMMENTS

Control Variates and Regression Estimators

The name "regression estimator" is adopted from the literatures of sampling survey (e.g., Cochran, 1977, Chapter 7) and Monte Carlo integration (e.g., Hammersley and Handscomb, 1964), and should be distinguished from "regression estimation" described by KS (Section 2.3). Specifically, the idea is to exploit the fact that if the PS model is correct, then $\hat{\eta}$ asymptotically has mean μ_1 (to be estimated) and $\hat{\xi}$ mean 0 (known). That is, $\hat{\xi}$ serve as auxiliary variables (in the terminology of survey sampling) or control variates (in that of Monte Carlo integration). Variance reduction can be achieved by using $\tilde{E}(\hat{\eta}) - b\tilde{E}(\hat{\xi})$, instead of $\hat{\mu}_{IPW} = \tilde{E}(\hat{\eta})$, with b an estimated regression coefficient of $\hat{\eta}$ against $\hat{\xi}$.

The control variates for $\tilde{\mu}_{REG}$ in Section 2 include $(\hat{\pi}^{-1}T - 1)\hat{m}_1$ and $(T - \hat{\pi})[\hat{\pi}(1-\hat{\pi})]^{-1}\partial\hat{\pi}/\partial\gamma$, the second of which is the score function for the PS model and is necessary for asymptotic optimality in Proposition 4(ii). If the PS model is correct, then $\tilde{\mu}_{REG}$ is always at least as efficient as $\hat{\mu}_{IPW}$ in the raw version, that is, $\hat{\mu}_{AIPW}(0)$, but not always than $\hat{\mu}_{IPW}$ in the ratio version. However, the indefiniteness can be easily resolved. If the control variate $\hat{\pi}^{-1}T - 1$ is added, or $(1, \hat{m}_1)^\top$ substituted for $\hat{m}_1$, then $\tilde{\mu}_{REG}$ always gains efficiency over both versions of $\hat{\mu}_{IPW}$. Furthermore, if $(1, h, \hat{m}_1)^\top$ is substituted for $\hat{m}_1$, then $\tilde{\mu}_{REG}$ always gains efficiency also over the estimator $\hat{\mu}_{AIPW}(h)$.

Causal Inference

Causal inference involves estimation of both μ_1 and μ_0. Similar estimators of μ_0 can be separately defined by replacing T, $\hat{\pi}$ and $\hat{m}_1$ with $1 - T$, $1 - \hat{\pi}$ and $\hat{m}_0$, where $m_0 = E(Y|T = 0, X)$. The control variates $((1-\hat{\pi})^{-1}(1-T) - 1)(1, \hat{m}_0)^\top$ for estimating μ_0 differ from $(\hat{\pi}^{-1}T - 1)(1, \hat{m}_1)^\top$ for estimating μ_1. As a consequence, even though $\tilde{\mu}_{1,REG}$ or $\tilde{\mu}_{0,REG}$ individually gains efficiency over $\hat{\mu}_{1,IPW}$ or $\hat{\mu}_{0,IPW}$, the difference $\tilde{\mu}_{1,REG} - \tilde{\mu}_{0,REG}$ does not necessarily gain efficiency over $\hat{\mu}_{1,IPW} - \hat{\mu}_{0,IPW}$. The problem can be overcome by using a combined set of control variates, say, $[\hat{\pi}^{-1}T - (1-\hat{\pi})^{-1}(1-T)](\hat{\pi}, 1-\hat{\pi}, \hat{\pi}\hat{m}_0, (1-\hat{\pi})\hat{m}_1)^\top$. Then $\tilde{\mu}_{1,REG} - \tilde{\mu}_{0,REG}$ maintains optimality in using control variates in the sense of Proposition 4(ii), in addition to local efficiency and double robustness. The mechanism of using a common set of control variates for estimating both μ_1 and μ_0 is automatic in the likelihood PS approach of Tan (2006).

PS Stratification

KS (Section 2.2) described the stratification estimator of Rosenbaum and Rubin (1983) as a way "to coarsen the estimated propensity score into a few categories and compute weighted averages of the mean response across categories." It is helpful to rewrite the estimator in their display (6) as

$$\hat{\mu}_{\text{strat}} = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i Y_i}{\hat{\pi}_{\text{strat}}(X_i)},$$

where $\hat{\pi}_{\text{strat}}(X) = \sum_{i=1}^{n} T_i 1\{\hat{\pi}(X_i) \in S_j\} / \sum_{i=1}^{n} 1\{\hat{\pi}(X_i) \in \hat{S}_j\}$ if $\hat{\pi}(X) \in \hat{S}_j$ (the jth estimated PS stratum), $j = 1, \ldots, s$. That is, $\hat{\mu}_{\text{strat}}$ is exactly an IPW estimator based on the discretized $\hat{\pi}_{\text{strat}}(X)$. Comparison between $\hat{\mu}_{\text{strat}}$ and $\hat{\mu}_{IPW}$ is subject to the usual bias-variance trade-off. On one hand, $\hat{\mu}_{\text{strat}}$ often has smaller variance than $\hat{\mu}_{IPW}$. On the other hand, the asymptotic limit of $\hat{\mu}_{\text{strat}}$ can be shown to be

$$\sum_{j=1}^{s} \frac{E[\pi(X)m_1(X)|\pi^*(X) \in S_j^*]}{E[\pi(X)|\pi^*(X) \in S_j^*]} P\big(\pi^*(X) \in S_j^*\big),$$

where $\pi^*(X)$ is the limit of $\hat{\pi}(X)$, which agrees with the true $\pi(X)$ if the PS model is correct, and S_j^* is that of $\hat{S}_j$. The ratio inside the above sum is the within-stratum average of $m_1(X)$ weighted proportionally to $\pi(X)$. Therefore, $\hat{\mu}_{\text{strat}}$ is inconsistent unless $\pi(X)$ or $m_1(X)$ is constant within each stratum (cf. KS's discussion about crude DR in Section 2.4). The asymptotic bias depends on the joint behavior of $m_1(X)$ and $\pi(X)$, and can be substantial if $m_1(X)$ varies where $\pi(X) \approx 0$ varies so that $m_1(X)$ are weighted differentially, say, by a factor of 10 at two X's with $\pi(X) = 0.01$ and 0.1.

Simulations

KS designed a simulation setup with an OR and a PS model appearing to be "nearly correct." The response is generated as $Y = 210 + 27.4Z_1 + 13.7Z_2 + 13.7Z_3 + 13.7Z_4 + \epsilon$, and the propensity score $\pi = \text{expit}(-Z_1 + 0.5Z_2 - 0.25Z_3 - 0.1Z_4)$, where ϵ and (Z_1, Z_2, Z_3, Z_4) are independent, standard normal. The covariates seen by the statistician are $X_1 = \exp(Z_1/2)$, $X_2 = Z_2/(1 + \exp(Z_1)) + 10$, $X_3 = (Z_1Z_3/25 + 0.6)^3$ and $X_4 = (Z_2 + Z_4 + 20)^2$. The OR model is the linear model of Y against X, and the PS model is the logistic model of T against X.

In the course of replicating their simulations, we accidentally discovered that the following models also appear to be "nearly correct." The covariates seen by the statistician are the same X_1, X_2, X_3, but $X_4 = (Z_3 + Z_4 + 20)^2$. The OR model is linear and the PS model is logistic as KS models. For one simulated dataset, Figures 1 and 2 present scatterplots and boxplots similar to Figures 2 and 3 in KS. For the OR model, the regression coefficients are highly significant and $R^2 = 0.97$. The correlation between the fitted values of Y under the correct and the misspecified OR models is 0.99, and that between the linear predictors under the correct and the misspecified PS models is 0.93. Tables 1 and 2 summarize our simulations for KS models and for the alternative models described above. The raw version of $\hat{\mu}_{IPW}$ is used. The estimators $\tilde{\mu}_{REG}^{(\text{m})}$ and $\hat{\mu}_{REG}^{(\text{m})}$ are defined as $\tilde{\mu}_{REG}$ and $\hat{\mu}_{REG}$ except that the score function for the PS model is dropped from $\hat{\xi}$. For these four estimators, $(1, \hat{m}_1)^\top$ is substituted for $\hat{m}_1$.

KS found that none of the DR estimators they tried improved upon the performance of the OR estimator; see also Table 1. This situation is consistent with the discussion in Section 2. The theory of DR estimation does *not* claim that a DR estimator is guaranteed to perform better than the OR estimator when the OR and the PS models are both misspecified, whether mildly or grossly. Therefore, KS's simulations serve as an example to remind us of this indefinite comparison.

On the other hand, neither is the OR estimator guaranteed to outperform DR estimators when the OR model is misspecified or even "nearly correct." As seen from Table 2, $\hat{\mu}_{OR}$ yields greater RMSE values than the DR estimators, $\hat{\mu}_{\text{WLS}}$, $\tilde{\mu}_{REG}$ and $\tilde{\mu}_{REG}^{(\text{m})}$ when the alternative, misspecified OR and PS models are both used. For $n = 200$, the bias of $\hat{\mu}_{OR}$ is 2.5 and that of $\tilde{\mu}_{REG}$ is 0.44, which differ substantially from the corresponding biases -0.56 and -1.8 in Table 1 when KS models are used.

The consequences of model misspecification are difficult to study, because the degree and direction of model misspecification are subtle, even elusive. For the dataset examined earlier, the absolute differences between the (highly correlated) fitted values of Y under the correct and the alternative, misspecified OR models present a more serious picture of model misspecification. In fact, the quartiles of these absolute differences are 2.0, 3.2 and 5.1, and the maximum is 20.

For both Tables 1 and 2, the DR estimators $\tilde{\mu}_{REG}$ and $\tilde{\mu}_{REG}^{(\text{m})}$ perform overall better than the other DR estimators $\hat{\mu}_{AIPW,\text{fix}}$ and $\hat{\mu}_{WLS}$. Compared with $\hat{\mu}_{WLS}$, $\tilde{\mu}_{REG}$ has MSE reduced by 15–20% (Table 1) or by 20–25% (Table 2) when the PS model is correct but the OR model is misspecified, which agrees with the optimality property of $\tilde{\mu}_{REG}$ in Proposition 4(ii). Even the simplified estimator $\tilde{\mu}_{REG}^{(\text{m})}$ gains similar efficiency, although the gain is not guaranteed in theory. The non-DR estimators $\hat{\mu}_{REG}$ and $\hat{\mu}_{REG}^{(\text{m})}$ sometimes have sizeable biases even when the PS model is correct.

Summary

One of the main points of KS is that two (moderately) misspecified models are not necessarily better than one. This point is valuable. But at the same time, neither are two misspecified models necessarily worse than one. Practitioners may choose to implement either of the OR and the PS-DR approaches, each with its own characteristics. It is helpful for statisticians to

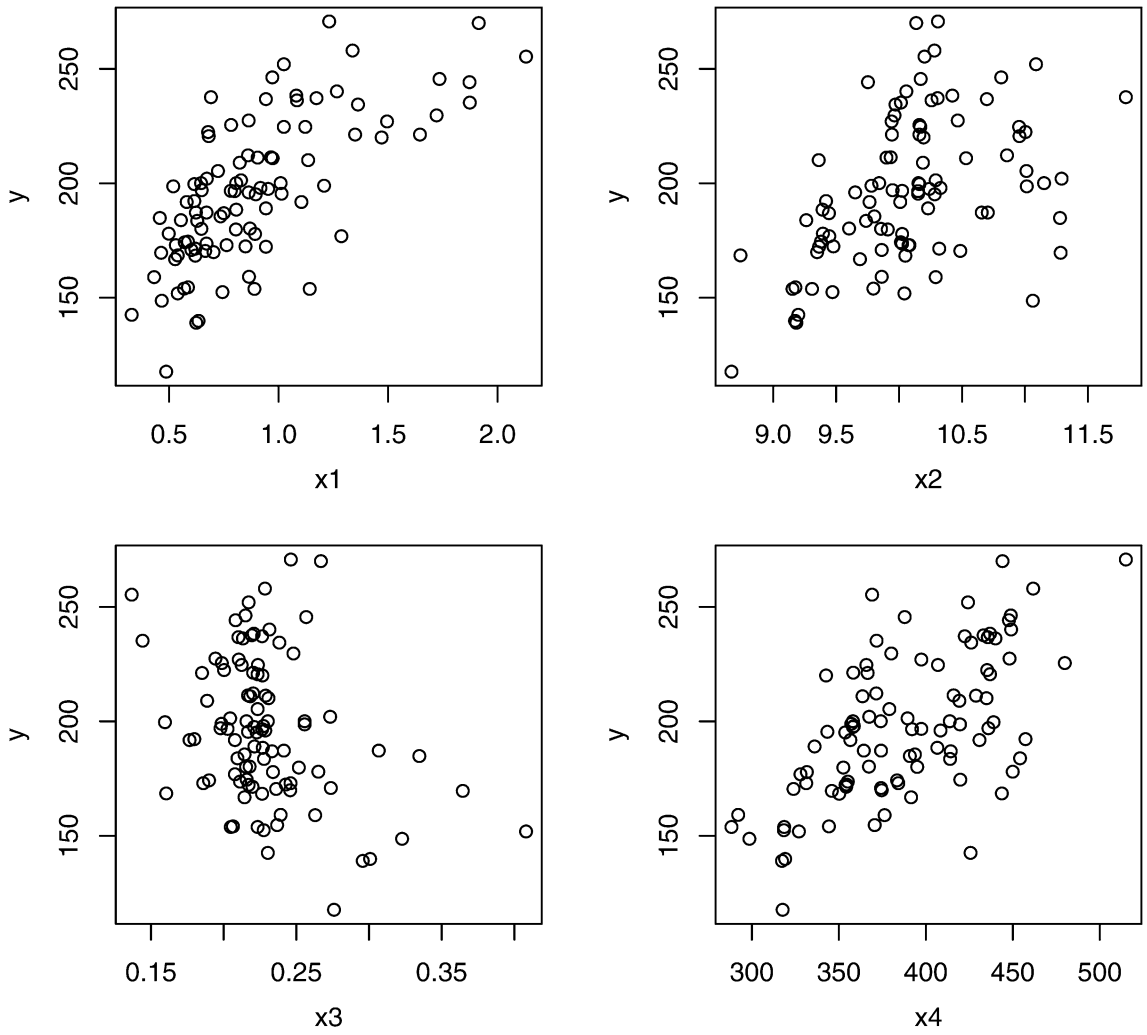


FIG. 1. *Scatterplots of response versus covariates* (*alternative models*).

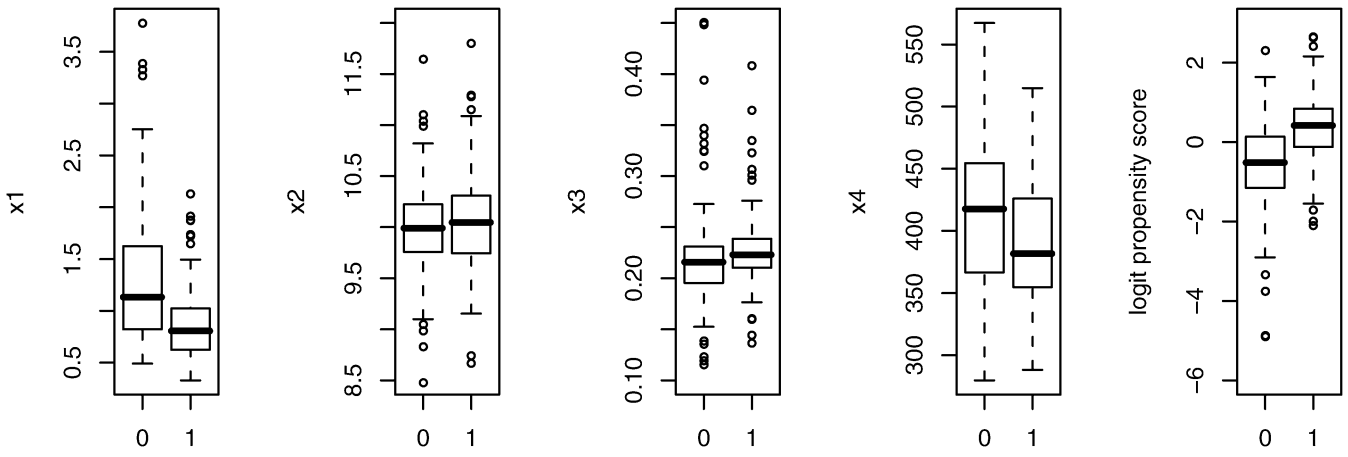


FIG. 2. *Boxplots of covariates and propensity scores* (*alternative models*).

TABLE 1
Numerical comparison of estimators of μ_1 (KS models)

	Method	Bias	% Bias	RMSE	MAE	Bias	% Bias	RMSE	MAE
$n = 200$			π-model correct				π-model incorrect		
	IPW	0.080	0.64	12.6	6.11	16	32	52.7	8.99
	strat	−1.1	−37	3.20	2.04	−2.9	−93	4.28	3.11
			y-model correct				y-model incorrect		
	OLS	−0.025	−0.99	2.47	1.68	−0.56	17	3.33	2.19
			y-model correct				y-model incorrect		
π-model	AIPW_{fix}	−0.024	−0.96	2.47	1.67	0.24	6.9	3.44	2.06
correct	WLS	−0.025	−1.0	2.47	1.68	0.39	13	2.99	1.89
	$\text{REG}_{\text{tilde}}$	−0.025	−1.0	2.47	1.69	0.14	5.2	2.73	1.76
	REG_{hat}	−0.52	−20	2.63	1.73	−0.52	−19	2.81	1.78
	$\text{REG}^{(\text{m})}_{\text{tilde}}$	−0.024	−0.98	2.47	1.68	0.24	8.9	2.74	1.79
	$\text{REG}^{(\text{m})}_{\text{hat}}$	−0.21	−8.4	2.48	1.68	−0.086	−3.2	2.65	1.74
π-model	AIPW_{fix}	−0.026	−1.0	2.48	1.71	−5.1	−44	12.6	3.75
incorrect	WLS	−0.026	−1.0	2.47	1.70	−2.2	−69	3.91	2.77
	$\text{REG}_{\text{tilde}}$	−0.027	−1.1	2.47	1.71	−1.8	−62	3.47	2.41
	REG_{hat}	−0.45	−18	2.60	1.71	−2.2	−76	3.68	2.53
	$\text{REG}^{(\text{m})}_{\text{tilde}}$	−0.026	−1.1	2.47	1.69	−2.0	−68	3.56	2.47
	$\text{REG}^{(\text{m})}_{\text{hat}}$	−0.13	−5.3	2.48	1.68	−2.2	−77	3.68	2.59
$n = 1000$			π-model correct				π-model incorrect		
	IPW	0.098	2.0	4.98	3.04	68	9.2	746	14.7
	strat	−1.1	−86	1.71	1.24	−2.9	−214	3.22	2.94
			y-model correct				y-model incorrect		
	OLS	−0.047	−4.0	1.15	0.770	−0.85	−56	1.75	1.15
			y-model correct				y-model incorrect		
π-model	AIPW_{fix}	−0.046	−4.0	1.15	0.766	0.043	2.6	1.63	1.11
correct	WLS	−0.046	−4.0	1.15	0.769	0.12	8.7	1.37	0.943
	$\text{REG}_{\text{tilde}}$	−0.046	−4.0	1.15	0.773	0.048	3.9	1.23	0.809
	REG_{hat}	−0.13	−11	1.16	0.796	−0.077	−6.3	1.23	0.812
	$\text{REG}^{(\text{m})}_{\text{tilde}}$	−0.046	−4.0	1.15	0.770	0.092	7.3	1.26	0.870
	$\text{REG}^{(\text{m})}_{\text{hat}}$	−0.083	−7.2	1.15	0.768	0.024	1.9	1.24	0.857
π-model	AIPW_{fix}	−0.10	−6.5	1.61	0.769	−26	−8.5	308	5.56
incorrect	WLS	−0.048	−4.1	1.15	0.764	−3.0	−203	3.38	3.05
	$\text{REG}_{\text{tilde}}$	−0.046	−4.0	1.15	0.764	−1.7	−120	2.21	1.73
	REG_{hat}	−0.045	−3.9	1.16	0.786	−1.7	−122	2.24	1.75
	$\text{REG}^{(\text{m})}_{\text{tilde}}$	−0.046	−4.0	1.15	0.763	−2.1	−152	2.48	2.04
	$\text{REG}^{(\text{m})}_{\text{hat}}$	−0.058	−5.0	1.16	0.771	−2.2	−158	2.57	2.15

promote a common, rigorous understanding of each approach and to investigate new ways for improvement. We welcome KS's article and the discussion as a step forward in this direction.

ACKNOWLEDGMENTS

We thank Xiao-Li Meng and Dylan Small for helpful comments.

REFERENCES

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley New York. MR0474575

HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66** 315–331. MR1612242

HAMMERSLEY, J. M. and HANDSCOMB, D. C. (1964). *Monte Carlo Methods*. Methuen, London. MR0223065

TABLE 2
Numerical comparison of estimators of μ_1 (alternative models)

	Method	Bias	% Bias	RMSE	MAE	Bias	% Bias	RMSE	MAE
$n = 200$			π-model correct				π-model incorrect		
	IPW	0.080	0.64	12.6	6.11	18	34	55.7	9.61
	strat	−1.1	−37	3.20	2.04	−1.1	−36	3.22	2.21
			y-model correct				y-model incorrect		
	OLS	−0.025	−0.99	2.47	1.68	2.5	80	4.04	2.73
			y-model correct				y-model incorrect		
π-model	AIPW_{fix}	−0.024	−0.96	2.47	1.67	0.53	14	3.82	2.32
correct	WLS	−0.025	−1.0	2.47	1.68	0.83	28	3.09	1.96
	$\text{REG}_{\text{tilde}}$	−0.025	−1.0	2.47	1.69	0.33	13	2.63	1.71
	REG_{hat}	−0.52	−20	2.63	1.73	−0.34	−13	2.70	1.74
	$\text{REG}^{(\text{m})}_{\text{tilde}}$	−0.024	−0.98	2.47	1.68	0.45	17	2.74	1.78
	$\text{REG}^{(\text{m})}_{\text{hat}}$	−0.21	−8.4	2.48	1.68	0.09	3.6	2.63	1.74
π-model	AIPW_{fix}	−0.024	−0.97	2.48	1.71	−2.5	−21	12.2	2.72
incorrect	WLS	−0.026	−0.10	2.47	1.70	0.33	11	3.11	2.05
	$\text{REG}_{\text{tilde}}$	−0.025	−0.10	2.47	1.71	0.44	16	2.74	1.80
	REG_{hat}	−0.42	−17	2.56	1.71	−0.026	−0.95	2.74	1.78
	$\text{REG}^{(\text{m})}_{\text{tilde}}$	−0.025	−1.0	2.47	1.69	0.31	11	2.83	1.80
	$\text{REG}^{(\text{m})}_{\text{hat}}$	−0.22	−8.9	2.48	1.71	0.035	1.3	2.76	1.77
$n = 1000$			π-model correct				π-model incorrect		
	IPW	0.098	2.0	4.98	3.04	80	8.5	951	16.8
	strat	−1.1	−86	1.71	1.24	−0.96	−72	1.65	1.17
			y-model correct				y-model incorrect		
	OLS	−0.047	−4.0	1.15	0.770	2.2	152	2.67	2.21
			y-model correct				y-model incorrect		
π-model	AIPW_{fix}	−0.046	−4.0	1.15	0.766	0.061	3.3	1.87	1.17
correct	WLS	−0.046	−4.0	1.15	0.769	0.22	16	1.39	0.957
	$\text{REG}_{\text{tilde}}$	−0.046	−4.0	1.15	0.773	0.12	10	1.21	0.818
	REG_{hat}	−0.13	−11	1.16	0.796	−0.012	−0.97	1.19	0.801
	$\text{REG}^{(\text{m})}_{\text{tilde}}$	−0.046	−4.0	1.15	0.770	0.14	12	1.25	0.849
	$\text{REG}^{(\text{m})}_{\text{hat}}$	−0.083	−7.2	1.15	0.768	0.069	5.7	1.22	0.826
π-model	AIPW_{fix}	−0.12	−6.3	1.83	0.780	−31	−6.9	441	2.92
incorrect	WLS	−0.048	−4.1	1.15	0.768	−0.55	−38	1.55	1.12
	$\text{REG}_{\text{tilde}}$	−0.044	−3.9	1.15	0.765	0.61	46	1.46	0.946
	REG_{hat}	−0.099	−8.5	1.16	0.787	0.57	43	1.45	0.910
	$\text{REG}^{(\text{m})}_{\text{tilde}}$	−0.045	−3.9	1.15	0.757	0.22	17	1.29	0.847
	$\text{REG}^{(\text{m})}_{\text{hat}}$	−0.16	−14	1.17	0.764	0.13	10	1.28	0.836

ROBINS, J. M. and ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* **90** 122–129. MR1325119

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90** 106–121. MR1325118

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

TAN, Z. (2006). A distributional approach for causal inference using propensity scores. *J. Amer. Statist. Assoc.* **101** 1619–1637. MR2279484

Statistical Science
2007, Vol. 22, No. 4, 569–573
DOI: 10.1214/07-STS227B
Main article DOI: 10.1214/07-STS227
© Institute of Mathematical Statistics, 2007

Comment: Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data

Anastasios A. Tsiatis and Marie Davidian

INTRODUCTION

We congratulate Drs. Kang and Schafer (KS henceforth) for a careful and thought-provoking contribution to the literature regarding the so-called "double robustness" property, a topic that still engenders some confusion and disagreement. The authors' approach of focusing on the simplest situation of estimation of the population mean μ of a response y when y is not observed on all subjects according to a missing at random (MAR) mechanism (equivalently, estimation of the mean of a potential outcome in a causal model under the assumption of no unmeasured confounders) is commendable, as the fundamental issues can be explored without the distractions of the messier notation and considerations required in more complicated settings. Indeed, as the article demonstrates, this simple setting is sufficient to highlight a number of key points.

As noted eloquently by Molenberghs (2005), in regard to how such missing data/causal inference problems are best addressed, two "schools" may be identified: the "likelihood-oriented" school and the "weighting-based" school. As we have emphasized previously (Davidian, Tsiatis and Leon, 2005), we prefer to view inference from the vantage point of semiparametric theory, focusing on the assumptions embedded in the statistical models leading to different "types" of estimators (i.e., "likelihood-oriented" or "weighting-based") rather than on the forms of the estimators themselves. In this discussion, we hope to complement the presentation of the authors by elaborating on this point of view.

Anastasios A. Tsiatis is Drexel Professor of Statistics at North Carolina State University, Raleigh, North Carolina 27695-8203, USA (e-mail: tsiatis@stat.ncsu.edu). Marie Davidian is William Neal Reynolds Professor of Statistics at North Carolina State University, Raleigh, North Carolina 27695-8203, USA (e-mail: davidian@stat.ncsu.edu).

Throughout, we use the same notation as in the paper.

SEMIPARAMETRIC THEORY PERSPECTIVE

As demonstrated by Robins, Rotnitzky and Zhao (1994) and Tsiatis (2006), exploiting the relationship between so-called *influence functions* and estimators is a fruitful approach to studying and contrasting the (large-sample) properties of estimators for parameters of interest in a statistical model. We remind the reader that a *statistical model* is a class of densities that could have generated the observed data. Our presentation here is for scalar parameters such as μ, but generalizes readily to vector-valued parameters. If one restricts attention to estimators that are *regular* (i.e., not "pathological"; see Davidian, Tsiatis and Leon, 2005, page 263 and Tsiatis 2006, pages 26–27), then, for a parameter μ in a parametric or semiparametric statistical model, an estimator $\widehat{\mu}$ for μ based on independent and identically distributed observed data z_i, $i = 1, \ldots, n$, is said to be *asymptotically linear* if it satisfies

$$n^{1/2}(\widehat{\mu} - \mu_0) = n^{-1/2} \sum_{i=1}^{n} \varphi(z_i) + o_p(1) \tag{1}$$

for $\varphi(z)$ with $E\{\varphi(z)\} = 0$ and $E\{\varphi^2(z)\} < \infty$, where μ_0 is the true value of μ generating the data, and expectation is with respect to the true distribution of z. The function $\varphi(z)$ is the *influence function* of the estimator $\widehat{\mu}$. A regular, asymptotically linear estimator with influence function $\varphi(z)$ is consistent and asymptotically normal with asymptotic variance $E\{\varphi^2(z)\}$. Thus, there is an inextricable connection between estimators and influence functions in that the asymptotic behavior of an estimator is fully determined by its influence function, so that it suffices to focus on the influence function when discussing an estimator's

properties. Many of the estimators discussed by KS are regular and asymptotically linear; in the sequel, we refer to regular and asymptotically linear estimators as simply "estimators."

We capitalize on this connection by considering the problem of estimating μ in the setting in KS in terms of statistical models that may be assumed for the observed data, from which influence functions corresponding to estimators valid under the assumed models may be derived. In the situation studied by KS, the "full" data that would ideally be observed are (t, x, y); however, as y is unobserved for some subjects, the observed data available for analysis are $z = (t, x, ty)$. As noted by KS, the MAR assumption states that y and t are conditionally independent given x; for example, $P(t = 1|y, x) = P(t = 1|x)$. Under this assumption, all joint densities for the observed data have the form

$$p(z) = p(y|x)^{I(t=1)} p(t|x) p(x), \tag{2}$$

where $p(y|x)$ is the density of y given x, $p(t|x)$ is the density of t given x, and $p(x)$ is the marginal density of x. Let $p_0(z)$ be the density in the class of densities of form (2) generating the observed data (the true joint density).

One may posit different statistical models by making different assumptions on the components of (2). We focus on three such models:

I. Make no assumptions on the forms of $p(x)$ or $p(t|x)$, leaving these entirely unspecified. Make a specific assumption on $p(y|x)$, namely, that $E(y|x) = m(x, \beta)$ for some given function $m(x, \beta)$ depending on parameters β $(p \times 1)$. Denote the class of densities satisfying these assumptions as $\mathcal{M}_I$.

II. Make no assumptions on the forms of $p(x)$ or $p(y|x)$. Make a specific assumption on $p(t|x)$ that $P(t = 1|x) = E(t|x) = \pi(x, \alpha)$ for some given function $\pi(x, \alpha)$ depending on parameters α $(s \times 1)$. Here, we also require the assumption that $P(t = 1|x) \geq \varepsilon > 0$ for all x and some ε. Denote the class of densities satisfying these assumptions as $\mathcal{M}_{II}$.

III. Make no assumptions on the form of $p(x)$, but make specific assumptions on $p(y|x)$ and $p(t|x)$, namely, that $E(y|x) = m(x, \beta)$ and $P(t = 1|x) = E(t|x) = \pi(x, \alpha) \geq \varepsilon > 0$ for all x and some ε for given functions $m(x, \beta)$ and $\pi(x, \alpha)$ depending on parameters β and α. The class of densities satisfying these assumptions is $\mathcal{M}_I \cap \mathcal{M}_{II}$.

All of I–III are semiparametric statistical models in that some aspects of $p(z)$ are left unspecified. Denote by $m_0(x)$ the true function $E(y|x)$ and by $\pi_0(x)$ the true function $P(t = 1|x) = E(t|x)$ corresponding to the true density $p_0(z)$.

Semiparametric theory yields the form of all influence functions corresponding to estimators for μ under each of the statistical models I–III. As discussed in Tsiatis (2006, page 52), loosely speaking, a consistent and asymptotically normal estimator for μ in a statistical model has the property that, for all $p(z)$ in the class of densities defined by the model, $n^{1/2}(\widehat{\mu} - \mu) \overset{\mathcal{D}(p)}{\to} \mathcal{N}\{0, \sigma^2(p)\}$, where $\overset{\mathcal{D}(p)}{\to}$ means convergence in distribution under the density $p(z)$, and $\sigma^2(p)$ is the asymptotic variance of $\widehat{\mu}$ under $p(z)$.

If model I is correct, then $m_0(x) = m(x, \beta)$ for some β, and it may be shown (e.g., Tsiatis, 2006, Section 4.5) that all estimators for μ have influence functions of the form

$$m_0(x) - \mu + ta(x)\{y - m_0(x)\} \tag{3}$$

for arbitrary functions $a(x)$ of x. If model II is correct, then $\pi_0(x) = \pi(x, \alpha)$ for some α, and all estimators for μ have influence functions of the form

$$\frac{ty}{\pi_0(x)} + \frac{t - \pi_0(x)}{\pi_0(x)} h(x) - \mu \tag{4}$$

for arbitrary $h(x)$, which is well known from Robins, Rotnitzky and Zhao (1994). If model III is correct, then $m_0(x) = m(x, \beta)$ and $\pi_0(x) = \pi(x, \alpha)$ for some β and α, and influence functions for estimators $\widehat{\mu}$ have the form

$$\begin{aligned} &m_0(x) - \mu + ta(x)\{y - m_0(x)\} \\ &\quad + \frac{t - \pi_0(x)}{\pi_0(x)} h(x) \end{aligned} \tag{5}$$

for arbitrary $a(x)$ and $h(x)$. Depending on forms of $m(x, \beta)$ as a function of β and $\pi(x, \alpha)$ as a function of α, there will be restrictions on the forms of $a(x)$ and $h(x)$; see below.

We now consider estimators discussed by KS from the perspective of influence functions. The regression estimator $\widehat{\mu}_{OLS}$ in (7) of KS comes about naturally if one assumes model I is correct. In terms of influence functions, $\widehat{\mu}_{OLS}$ may be motivated by considering the influence function (3) with $a(x) = 0$, as this leads to the estimator $n^{-1} \sum_{i=1}^{n} m(x_i, \beta)$. In fact, although KS do not discuss it, the "imputation estimator" $\widehat{\mu}_{IMP} = n^{-1} \sum_{i=1}^{n} \{t_i y_i + (1 - t_i) m(x_i, \beta)\}$ may be motivated by taking $a(x) = 1$ in (3). Of course,

in practice, β must be estimated. In general, (3) implies that all estimators for μ that are consistent and asymptotically normal if model I is correct must be asymptotically equivalent to an estimator of the form

$$n^{-1}\sum_{i=1}^{n}[m(x_i,\widehat{\beta})+t_i\widetilde{a}(x_i)\{y_i-m(x_i,\widehat{\beta})\}], \tag{6}$$

where β is estimated by solving an estimating equation $\sum_{i=1}^{n} t_i A(x_i,\beta)\{y_i - m(x_i,\beta)\} = 0$ for $A(x,\beta)$ $(p \times 1)$. Because β is estimated, the influence function of the estimator (6) with a particular $\widetilde{a}(x)$ will not be exactly equal to (3) with $a(x) = \widetilde{a}(x)$; instead, it may be shown that the influence function of (6) is of form (3) with $a(x)$ in (3) equal to

$$\begin{aligned}&\widetilde{a}(x) - E[\{\pi_0(x)\widetilde{a}(x)-1\}m_\beta^T(x,\beta_0)]\\ &\quad\cdot[E\{\pi_0(x)A(x,\beta_0)m_\beta^T(x,\beta_0)\}]^{-1}\\ &\quad\cdot A(x,\beta_0),\end{aligned} \tag{7}$$

where $m_\beta(x,\beta)$ is the vector of partial derivatives of elements of $m(x,\beta)$ with respect to β, and β_0 is such that $m_0(x) = m(x,\beta_0)$.

The IPW estimator $\widehat{\mu}_{IPW\text{-}POP}$ in (3) of KS and its variants arise if one assumes model II. In particular, $\widehat{\mu}_{IPW\text{-}POP}$ can be motivated via the influence function (4) with $h(x) = -\mu$. The estimator $\widehat{\mu}_{IPW\text{-}NR}$ in (4) of KS follows from (4) with $h(x) = -E[y\{1-\pi(x)\}]/E[\{1-\pi(x)\}]$. In fact, if one restricts $h(x)$ in (4) to be a constant, then, using the fact that the expectation of the square of (4) is the asymptotic variance of the estimator, one may find the "best" such constant minimizing the variance as $h(x) = -E[y\{1-\pi(x)\}/\pi(x)]/E[\{1-\pi(x)\}/\pi(x)]$. An estimator based on this idea was given in (10) of Lunceford and Davidian (2004, page 2943). In general, as for model I, (4) implies that all estimators for μ that are consistent and asymptotically normal if model II is correct must be asymptotically equivalent to an estimator of the form

$$n^{-1}\sum_{i=1}^{n}\left\{\frac{t_i y_i}{\pi(x_i,\widehat{\alpha})}+\frac{t_i-\pi(x_i,\widehat{\alpha})}{\pi(x_i,\widehat{\alpha})}\widetilde{h}(x_i)\right\}, \tag{8}$$

where $\widehat{\alpha}$ is estimated by solving an equation of the form $\sum_{i=1}^{n}\{t_i-\pi(x_i,\alpha)\}B(x_i,\alpha)=0$ for some $(s\times 1)$ $B(x_i,\alpha)$, almost always maximum likelihood for binary regression. As above, because α is estimated, the influence function of (8) is equal to (4) with $h(x)$ equal to

$$\begin{aligned}&\widetilde{h}(x) - E[\pi_\alpha^T(x,\alpha_0)\{m_0(x)+\widetilde{h}(x)\}/\pi_0(x)]\\ &\quad\cdot[E\{B(x,\alpha_0)\pi_\alpha^T(x,\alpha_0)\}]^{-1}\\ &\quad\cdot B(x,\alpha_0)\pi_0(x),\end{aligned} \tag{9}$$

where $\pi_\alpha(x,\alpha)$ is the vector of partial derivatives of elements of $\pi(x,\alpha)$ with respect to α, and α_0 satisfies $\pi_0(x) = \pi(x,\alpha_0)$.

Doubly robust (DR) estimators are estimators that are consistent and asymptotically normal for models in $\mathcal{M}_I \cup \mathcal{M}_{II}$, that is, under the assumptions of model I or model II. When the true density $p_0(z) \in \mathcal{M}_I \cap \mathcal{M}_{II}$, then the influence function of any such DR estimator must be equal to (3) with $a(x) = 1/\pi_0(x)$ or, equivalently, equal to (4) with $h(x) = -m_0(x)$. Accordingly, when $p_0(z) \in \mathcal{M}_I \cap \mathcal{M}_{II}$, that is, both models have been specified correctly, all such DR estimators will have the same asymptotic variance. This also implies that, if both models are correctly specified, the asymptotic properties of the estimator do not depend on the methods used to estimate β and α.

KS discuss strategies for constructing DR estimators, and they present several specific examples: $\widehat{\mu}_{BC\text{-}OLS}$ in their equation (8); the estimators below (8) using POP or NR weights, which we denote as $\widehat{\mu}_{BC\text{-}POP}$ and $\widehat{\mu}_{BC\text{-}NR}$, respectively; the estimator $\widehat{\mu}_{WLS}$ in their equation (10); $\widehat{\mu}_{\pi\text{-}cov}$ in their equation (12); and a version of $\widehat{\mu}_{\pi\text{-}cov}$ equal to the estimator proposed by Scharfstein, Rotnitzky and Robins (1999) and Bang and Robins (2005), which we denote as $\widehat{\mu}_{SRR}$. The results for these estimators under the "Correct-Correct" scenarios ($\mathcal{M}_I \cap \mathcal{M}_{II}$) in Tables 5–8 of KS are consistent with the asymptotic properties above. We note that $\widehat{\mu}_{\pi\text{-}cov}$ is not DR under $\mathcal{M}_I \cup \mathcal{M}_{II}$ because of the additional assumption that the mean of y given π must be equal to a linear combination of basis functions in π. Making this additional assumption may not be unreasonable in practice; however, strictly speaking, it takes $\widehat{\mu}_{\pi\text{-}cov}$ outside the class of DR estimators discussed here, and hence we do not consider it in the remainder of this section. However, $\widehat{\mu}_{SRR}$ is still in this class.

KS suggest that a characteristic distinguishing the performance of DR estimators is whether or not the estimator is within or outside the augmented inverse-probability weighted (AIPW) class. We find this distinction artificial, as all of the above estimators $\widehat{\mu}_{BC\text{-}OLS}$, $\widehat{\mu}_{BC\text{-}POP}$, $\widehat{\mu}_{BC\text{-}NR}$, $\widehat{\mu}_{WLS}$ and $\widehat{\mu}_{SRR}$ can be expressed in an AIPW form. Namely, all of these estimators are algebraically exactly of the form (8)

with $\widetilde{h}(x_i)$ replaced by a term $-\widehat{\gamma} - m(x_i, \widehat{\beta})$, where $\widehat{\gamma}_{BC\text{-}OLS} = \widehat{\gamma}_{WLS} = \widehat{\gamma}_{SRR} = 0$,

$$\begin{aligned} &\widehat{\gamma}_{BC\text{-}POP} \\ &\quad = \frac{n^{-1}\sum_{i=1}^{n}(t_i/\widehat{\pi}_i)(y_i - \widehat{m}_i)}{n^{-1}\sum_{i=1}^{n} t_i/\widehat{\pi}_i} \quad \text{and} \\ &\widehat{\gamma}_{BC\text{-}NR} \\ &\quad = \frac{n^{-1}\sum_{i=1}^{n}(t_i(1-\widehat{\pi}_i)/\widehat{\pi}_i)(y_i - \widehat{m}_i)}{n^{-1}\sum_{i=1}^{n} t_i(1-\widehat{\pi}_i)/\widehat{\pi}_i}, \end{aligned} \tag{10}$$

where we write $\widehat{\pi}_i = \pi(x_i, \widehat{\alpha})$ and $\widehat{m}_i = m(x_i, \widehat{\beta})$ for brevity. For $\widehat{\mu}_{WLS}$ and $\widehat{\mu}_{SRR}$, this identity follows from the fact that $\sum_{i=1}^{n} \frac{t_i}{\widehat{\pi}_i}(y_i - \widehat{m}_i) = 0$, which for $\widehat{\mu}_{WLS}$ holds because KS restrict to $m(x, \beta) = x^T\beta$, with x including a constant term. Thus, we contend that issues of performance under $\mathcal{M}_I \cup \mathcal{M}_{II}$ are not linked to whether or not a DR estimator is AIPW, but, rather, are a consequence of forms of the influence functions of estimators under $\mathcal{M}_I$ or $\mathcal{M}_{II}$. In particular, under model II, it follows that the above estimators have influence functions of the form (4) with $h(x)$ equal to (9) with $\widetilde{h}(x) = -\{\gamma^* + m(x, \beta^*)\}$, where γ^* and β^* are the limits in probability of $\widehat{\gamma}$ and $\widehat{\beta}$, respectively. Thus, features determining performance of these estimators when model II is correct are how close $\gamma^* + m(x, \beta^*)$ is to $m_0(x)$ and how α is estimated, where maximum likelihood is the optimal choice. In fact, this perspective reveals that, for fixed $m(x, \beta)$, using ideas similar to those in Tan (2006), the optimal choice of $\widehat{\gamma}$ is as in $\widehat{\gamma}_{BC\text{-}NR}$ with $t_i(1-\widehat{\pi}_i)/\widehat{\pi}_i$ replaced by $t_i(1-\widehat{\pi}_i)/\widehat{\pi}_i^2$.

Similarly, under model I, the influence functions of these estimators are of the form (3) with $a(x)$ equal to (7) with $\widetilde{a}(x) = \psi_1/\pi(x, \alpha^*) + \psi_2$, where α^* is the limit in probability of $\widehat{\alpha}$ and $\psi_1 = 1$ and $\psi_2 = 0$ for $\widehat{\mu}_{BC\text{-}OLS}$, $\widehat{\mu}_{WLS}$ and $\widehat{\mu}_{SRR}$; $\psi_1 = 1/E\{\pi_0(x)/\pi(x, \alpha^*)\}$ and $\psi_2 = 0$ for $\widehat{\mu}_{BC\text{-}POP}$; and ψ_1 and ψ_2 for $\widehat{\mu}_{BC\text{-}NR}$ are more complicated expectations involving $\pi_0(x)$ and $\pi(x, \alpha^*)$. Thus, under model I, features determining performance of these estimators are the form of $\widetilde{a}(x)$ and how β is estimated through the choice of $A(x, \beta)$.

We may interpret some of the results in Tables 5, 6 and 8 of KS in light of these observations. Under the "π-model Correct–y-model Incorrect" scenario ($\mathcal{M}_{II} \cap \mathcal{M}_I^c$), $\widehat{\mu}_{BC\text{-}OLS}$, $\widehat{\mu}_{WLS}$ and $\widehat{\mu}_{SRR}$ show some nontrivial differences in performance, which, from above, are likely attributable to differences in $m(x, \beta^*)$. Under the "π-model Incorrect–y-model Correct" ($\mathcal{M}_I \cap \mathcal{M}_{II}^c$), all three estimators share the same $\widetilde{a}(x)$ but use different methods to estimate β, so that any differences are dictated entirely by the choice of $A(x, \beta)$. The poor performance of $\widehat{\mu}_{SRR}$ can be understood from this perspective: "β" for this estimator is actually β in the model $m(x, \beta)$ used by the other two estimators concatenated by an additional element, the coefficient of $\widehat{\pi}_i^{-1}$. The $A(x, \beta)$ for $\widehat{\mu}_{SRR}$ thus involves a design matrix that is unstable for small $\widehat{\pi}_i$, consistent with the comment of KS at the end of their Section 3.

In summary, we believe that studying the performance of estimators via their influence functions can provide useful insights. Our preceding remarks refer to large-sample performance, which depends directly on the influence function. Estimators with the same influence function can exhibit different finite-sample properties. It may be possible via higher-order expansions to gain an understanding of some of this behavior; to the best of our knowledge, this is an open question.

BOTH MODELS INCORRECT

The developments in the previous section are relevant in $\mathcal{M}_I \cup \mathcal{M}_{II}$. Key themes of KS are performance of DR and other estimators outside this class; that is, when both the models $\pi(x, \alpha)$ and $m(x, \beta)$ are incorrectly specified, and choice of estimator under these circumstances.

One way to study performance in this situation is through simulation. KS have devised a very interesting and instructive specific simulation scenario that highlights some important features of various estimators. In particular, the KS scenario emphasizes the difficulties encountered with some of the DR estimators when $\pi(x_i, \widehat{\alpha})$ is small for some x_i. Indeed, in our experience, poor performance of DR and IPW estimators in practice can result from few small $\pi(x_i, \widehat{\alpha})$. When there are small $\pi(x_i, \widehat{\alpha})$, as noted KS, responses are not observed for some portion of the x space. Consequently, estimators like $\widehat{\mu}_{OLS}$ rely on extrapolation into that part of the x space. KS have constructed a scenario where failure to observe y in a portion of the x space can wreak havoc on some estimators that make use of the $\pi(x_i, \widehat{\alpha})$ but has minimal impact on the quality of extrapolations for these x based on $m(x, \widehat{\beta})$. One could equally well build a scenario where the x for which y is unobserved are highly influential for the regression $m(x, \beta)$ and hence could result in deleterious performance of $\widehat{\mu}_{OLS}$. We thus reiterate the remark of KS that, although simulations can be illuminating, they cannot yield broadly applicable conclusions.

Given this, we offer some thoughts on other strategies for deriving estimators that may have some robustness properties under the foregoing conditions, that

is, offer good performance outside $\mathcal{M}_I \cup \mathcal{M}_{II}$. One approach may be to search outside the class of DR estimators valid under $\mathcal{M}_I \cup \mathcal{M}_{II}$. For example, as suggested by the simulations of KS, estimators in the spirit of $\widehat{\mu}_{\pi\text{-}cov}$, which impose additional assumptions rendering them DR in the strict sense only in a subset of $\mathcal{M}_I \cup \mathcal{M}_{II}$, may compensate for this restriction by yielding more robust performance outside $\mathcal{M}_I \cup \mathcal{M}_{II}$; further study along these lines would be interesting. An alternative tactic for searching outside $\mathcal{M}_I \cup \mathcal{M}_{II}$ may be to consider the form of influence functions (5) for estimators valid under $\mathcal{M}_I \cap \mathcal{M}_{II}$. For instance, a "hybrid" estimator of the form

$$n^{-1}\sum_{i=1}^{n}\Bigg[m(x_i,\widehat{\beta})I\{\pi(x_i,\widehat{\alpha})<\delta\} + \left\{\frac{t_i y_i}{\pi(x_i,\widehat{\alpha})}+\frac{t_i-\pi(x_i,\widehat{\alpha})}{\pi(x_i,\widehat{\alpha})}\widetilde{h}(x_i)\right\} \cdot I\{\pi(x_i,\widehat{\alpha})\geq\delta\}\Bigg],$$

for δ small, may take advantage of the desirable properties of both $\widehat{\mu}_{OLS}$ and DR estimators.

A second possible strategy for identifying robust estimators arises from the following observation. Consider the estimator

$$n^{-1}\sum_{i=1}^{n}\left\{\frac{t_i y_i}{\pi(x_i)}-\frac{t_i-\pi(x_i)}{\pi(x_i)}m(x_i,\widehat{\beta})\right\}. \tag{11}$$

If $\pi(x_i)=\pi(x_i,\widehat{\alpha})$, then (11) yields one form of a DR estimator. If $\pi(x_i)\equiv 1$, then (11) results in the imputation estimator. If $\pi(x_i)=\infty$, (11) reduces to $\widehat{\mu}_{OLS}$. This suggests that it may be possible to develop estimators based on alternative choices of $\pi(x_i)$ that may have good robustness properties. For example, a method for obtaining estimators $\pi(x_i,\widehat{\alpha})$ that shrinks these toward a common value may prove fruitful. The suggestion of KS to move away from logistic regression models for $\pi(x_i,\alpha)$ is in a similar spirit.

Finally, we note that yet another approach to developing estimators would be to start with the premise that one make no parametric assumption on the forms of $E(y|x)$ and $E(t|x)$ beyond some mild smoothness conditions. Here, it is likely that first-order asymptotic theory, as in the previous section, may no longer be applicable. It may be necessary to use higher-order asymptotic theory to make progress in this direction; see, for example, Robins and van der Vaart (2006).

CONCLUDING REMARKS

We again compliment the authors for their thoughtful and insightful article, and we appreciate the opportunity to offer our perspectives on this important problem. We look forward to new methodological developments that may overcome some of the challenges brought into focus by KS in their article.

ACKNOWLEDGMENT

This research was supported in part by Grants R01-CA051962, R01-CA085848 and R37-AI031789 from the National Institutes of Health.

REFERENCES

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189

DAVIDIAN, M., TSIATIS, A. A. and LEON, S. (2005). Semiparametric estimation of treatment effect in a pretest-posttest study without missing data. *Statist. Sci.* **20** 261–301. MR2189002

LUNCEFORD, J. K. and DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* **23** 2937–2960.

MOLENBERGHS, G. (2005). Discussion of "Semiparametric estimation of treatment effect in a pretest–posttest study with missing data," by M. Davidian, A. A. Tsiatis and S. Leon. *Statist. Sci.* **20** 289–292. MR2189002

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

ROBINS, J. and VAN DER VAART, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.* **34** 229–253. MR2275241

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Rejoinder to "Adjusting for nonignorable drop-out using semiparametric nonresponse models." *J. Amer. Statist. Assoc.* **94** 1135–1146. MR1731478

TAN, Z. (2006). A distributional approach for causal inference using propensity scores. *J. Amer. Statist. Assoc.* **101** 1619–1637. MR2279484

TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York. MR2233926

Statistical Science
2007, Vol. 22, No. 4, 574–580
DOI: 10.1214/07-STS227REJ
Main article DOI: 10.1214/07-STS227
© Institute of Mathematical Statistics, 2007

Rejoinder: Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data

Joseph D. Y. Kang and Joseph L. Schafer

1. CLARIFYING OUR POSITION ON DOUBLY ROBUST ESTIMATORS

We are grateful to the editors for eliciting comments from some of the most prominent researchers in this exciting and rapidly developing field. After we drafted our article, a number of important works on DR estimators appeared, including Tan's (2006) article on causal inference, the monograph by Tsiatis (2006) and the recent articles and technical reports cited by Robins, Sued, Lei-Gomez and Rotnitzky. The discussants' insightful remarks highlight these recent developments and bring us up to date.

Our purpose in writing this article was to provide unfamiliar readers with gentle introduction to DR estimators without the language of influence functions, using only simple concepts from regression analysis and survey inference. We wanted to show that DR estimators come in many different flavors. And, without minimizing the importance of the literature spawned by Robins, Rotnitzky and Zhao (1994), we wanted to draw attention to some older related methods from model-assisted survey sampling which arrive at a similar position from the opposite direction.

Despite the good performance of $\hat{\mu}_{OLS}$ in our simulated example, we have not and would not argue that it be used routinely and uncritically. The pitfalls of relying solely on outcome regression or y-modeling have been well documented for causal inference, where the rates of missing information are high and the impact of selection bias can be dramatic (e.g., Rubin, 2001). Nor do we wish to cast clouds of suspicion over all DR estimators in all circumstances. In many situations, they do work well. On the other hand, we still believe that procedures motivated by parametric y-models, when carefully implemented, remain a viable option and should not be categorically dismissed.

Under ignorability, the propensities $\pi_i = P(t_i = 1|x_i)$, $i = 1, \ldots, n$, play no role in likelihood-based or Bayesian inference about μ under a given y-model. If we had absolute faith in one parametric form for $P(y_i|x_i)$, then we could discard all information beyond the sufficient statistics for that model. But the propensities carry information that helps us evaluate the quality of the y-model, and we ignore this extra information at our peril, because no model is above criticism. No sensible statistician would argue that propensities should not be examined. But reasonable persons may differ over what role the propensities should play in formulating an estimator. Those who favor a semiparametric approach devise influence functions that combine inverse-propensity weights with regression predictions for y. Parametric modelers, on the other hand, may well argue that if the propensities reveal weaknesses in the y-model, then that model should be revised and corrected. The latter view has been expressed by Elliott and Little (2000) in the context of survey estimation, where the selection probabilities are known, but parallels to uncontrolled nonresponse and causal inference are obvious.

We believe that propensities are useful for model diagnosis and estimation, but we are still not convinced that they need to enter an influence function as inverse-propensity weights. The strength of weighting is that, if done properly, it protects an estimate from bias regardless of how y is distributed. But this strength can also be a weakness, because such a high level of protection is not always warranted. If the propensities are unrelated to the linear predictors from a good y-model, then

Joseph D. Y. Kang is Research Associate, The Methodology Center, 204 E. Calder Way, Suite 400, State College, Pennsylvania 16801, USA (e-mail: josephkang@stat.psu.edu). Joseph L. Schafer is Associate Professor, The Methodology Center, The Pennsylvania State University, 204 E. Calder Way, Suite 400, State College, Pennsylvania 16801, USA (e-mail: jls@stat.psu.edu).

weighting may be superfluous. If the propensities are poorly estimated or extreme, then combining weights with the regression predictions may do more harm than good. And if the propensities do reveal weaknesses in the y-model, inverse-propensity weights are not the only way to correct them.

2. RESPONSE TO TSIATIS AND DAVIDIAN

In their illuminating discussion, Tsiatis and Davidian demonstrate that a wide variety of estimators for μ can be expressed as the solution to an estimating equation based on an influence function. (One possible exception is the class of estimators based on propensity-score matching, which we have not discussed.) Influence functions present interesting results on semiparametric efficiency, but we find them appealing for other reasons as well. First, they show us how to compute a standard error for whatever estimator we choose. Second, they generalize nicely to finite-population sample surveys with complex designs. Regression techniques for complex surveys, as implemented in software packages like SUDAAN (Shah, Barnwell and Biler, 1997), are based on weighted influence functions, so any of the estimators described by Tsiatis and Davidian can be extended to surveys. Third, if we move on to causal inference, we must address the thorny issue of the inestimable partial correlation between the potential outcomes. Any estimator of an average causal effect makes a working assumption about this correlation (e.g., setting it to zero), but a standard error computed from an influence-function sandwich may still perform well when this working correlation is incorrect.

Tsiatis and Davidian mention that our estimator $\hat{\mu}_{\pi\text{-}cov}$, which incorporates propensity-related basis functions into the OLS procedure, is not consistent under $\mathcal{M}_I \cup \mathcal{M}_{II}$ unless the conditional mean of y_i happens to be a linear combination of the particular basis functions for π_i used in the OR model. This is certainly true for the usual asymptotic sequence in which the number of basis functions remains fixed as $n \to \infty$. But if we allow the basis to grow with the sample size (e.g., as in a smoothing spline), then it may become DR (Little and An, 2004). Given a large sample, a good data analyst will tend to fit a richer model than with a small sample. If the analyst is allowed to build a rich OR model that corrects for the kind of inadequacies shown in our Figure 4, then the OLS procedure based on the corrected OR model may be as good as any DR procedure.

We like the suggestion by Tsiatis and Davidian of using a hybrid estimator that combines inverse-propensity weighting for cases with moderate propensity and regression predictions for cases with small propensity, an idea echoed by van der Laan and Rubin (2006). As an alternative to a hard threshold δ at which the change is made, one could opt for a smoother transition by "weighting" each part of the influence function more or less depending on the estimated propensity. We also agree with Tsiatis and Davidian that estimators in the spirit of $\hat{\mu}_{\pi\text{-}cov}$ deserve more consideration even though they are not DR over $\mathcal{M}_I \cup \mathcal{M}_{II}$ in the usual asymptotic sequence. In the simulations of our article, we expressed m_i as a piecewise constant function of $\hat{\pi}_i$ with discontinuities at the sample quintiles of $\hat{\pi}_i$. Another version of $\hat{\mu}_{\pi\text{-}cov}$ that we have found to work well in many situations uses a linear spline in $\hat{\eta}_i = \log(\hat{\pi}_i/(1-\hat{\pi}_i))$ with knots at the quintiles.

3. RESPONSE TO TAN

Tan's important work on regression estimators connects the theory of influence functions to ideas of survey regression estimators and the use of control variates in importance sampling. His remarks and propositions are very helpful for understanding the behavior of IPW, OR and DR methods in realistic settings where all the models are incorrect.

We were initially puzzled by several of Tan's points but, upon further consideration, found them to be very insightful. He states that it is more constructive to view DR estimators as efficiency-enhanced versions of IPW than as bias-corrected versions of OR. We find both views helpful for understanding the nature and properties of DR methods. But, as he explains, there are theoretical reasons to expect that his carefully crafted DR estimators may lead to greater improvement over IPW than over a good OR model, because IPW is conservative whereas OR is aggressive.

We are still unsure why Tan states that IPW extrapolates explicitly whereas OR extrapolates implicitly. To us, fitting an OR model to respondents and using that model to predict for nonrespondents is a very obvious kind of extrapolation, especially if the leverage values for some nonrespondents are large relative to those of the respondents. But his points about extrapolation are well taken. All of our methods extrapolate. The assumption of ignorability is itself an extrapolation.

He also points out that estimating an average causal effect is more subtle than simply estimating the mean

of each potential outcome and taking the difference. This distinction is important in a semiparametric approach. A semiparametric method that is optimal for estimating two means may not be optimal for estimating the mean difference. Similarly, a method that is optimal for estimating a population average causal effect may not be optimal for estimating the average effect among the treated, or for estimating differences in average causal effects between subpopulations. As parametric assumptions about the OR model are discarded, it becomes important to tailor the estimation procedure to the estimand, which his regression estimators apparently do.

In Tan's simulations, his alternative model in which the analyst sees $X_4 = (Z_3 + Z_4 + 20)^2$ presents an interesting situation where OLS predicts the y_i's for the respondents almost perfectly ($R^2 \approx 0.99$), but the extrapolated linear predictions for the nonrespondents are biased because the unseen true values of y_i turn sharply away from those predictions in the region of low propensity. This is a perfect illustration of how the uncritical use of $\hat{\mu}_{OLS}$ can lead us astray. But in this example, propensity-based diagnostics reveal obvious deficiencies in the linear model. Taking the initial sample of $n = 200$ observations from our article, we fit the linear model to the respondents and a logistic propensity model to all cases given X_1, X_2, X_3, and Tan's alternative version of X_4. A plot of the observed residuals from the y-model versus the estimated logit-propensities from the π-model is shown in Figure 1. The loess curve clearly shows that the OLS predictions are biased in the region of high propensity (where it does not really matter) and in the region of low propensity (where it matters very much). If we account for this trend by introducing the squared linear predictor from the logit model $\hat{\eta}_i^2 = (x_i^T \hat{\alpha})^2$ as one more covariate in the y-model, the performance of $\hat{\mu}_{OLS}$ greatly improves. Even better performance is obtained with splines, which tend to predict better than ordinary polynomials over the whole range of $\hat{\eta}_i$'s. We created a linear spline basis for $\hat{\eta}_i$ with four knots located at the sample quintiles of $\hat{\eta}_i$. That is, we added the four covariates

$$(\hat{\eta}_i - k_1)_+, \quad (\hat{\eta}_i - k_2)_+, \quad (\hat{\eta}_i - k_3)_+, \quad (\hat{\eta}_i - k_4)_+ \tag{1}$$

to the y model, where $(z)_+ = \max(0, z)$ and k_1, k_2, k_3, k_4 are the knots. Over 1000 samples, we found that this new version of $\hat{\mu}_{OLS}$ (which, in our article, we would have called $\hat{\mu}_{\pi\text{-}cov}$) performed as well as any of Tan's estimators in the scenario where both models were incorrect. With $n = 200$, we obtained bias $= 0.16$, % bias $= 5.70$, RMSE $= 2.78$ and MAE $= 1.78$. With $n = 1000$, we obtained bias $= 0.30$, % bias $= 24.6$, RMSE $= 1.27$ and MAE $= 0.88$. The performance of Tan's regression estimators in these simulations is impressive. The performance of $\hat{\mu}_{OLS}$ is equally impressive if we allow the analyst to make a simple correction to adjust for the y-model's obvious lack of fit.

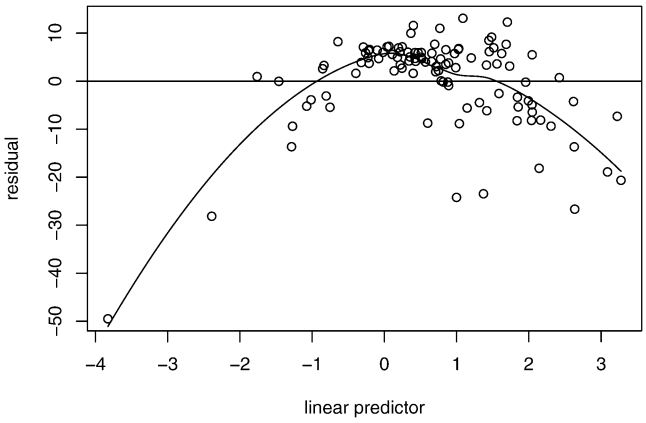


FIG. 1. *Scatterplot of raw residuals from linear y-model fit to respondents in Tan's alternative model, versus the linear predictors from a logistic π-model, with local polynomial (loess) fit.*

4. RESPONSE TO RIDGEWAY AND MCCAFFREY

Ridgeway and McCaffrey correctly observe that, for estimating propensity scores, there are many good alternatives to logistic regression. In addition to their work on the generalized boosted model (GBM), some have been estimating propensities using classification trees (Luellen, Shadish and Clark, 2005) and neural networks (King and Zeng, 2002).

A rich propensity model should improve the performance of the weighted estimator. The advantage of procedures like classification trees and GBM is that they allow us to search through a large space of π-models, accounting for the effects of many covariates and their interactions, thereby reducing bias in the resulting estimator regardless of how y_i is distributed. These procedures may also reduce variance, because, as explained by Tan, in a sequence of increasingly rich propensity models, the asymptotic variance of an augmented IPW estimator decreases to the semiparametric bound. In principle, one could apply similar procedures like regression trees to create a rich y-model. But, as Ridgeway and McCaffrey point out, this raises the possibility of data snooping. As we search through larger and more complicated spaces to find the best

y-model, it becomes increasingly difficult to compute honest standard errors.

Ridgeway and McCaffrey's simulations with the extra interaction term again reveal the dangers of uncritically relying on $\hat{\mu}_{OLS}$. This interaction increases the degree to which the additive and linear y-model is misspecified, so in this scenario we would expect the performance of $\hat{\mu}_{OLS}$ to worsen. The final columns of their Tables 1 and 2 show that, when this interaction is present, propensity-based and DR estimators strongly outperform $\hat{\mu}_{OLS}$. Using the wrong covariates in the propensity model does little harm to the flexible GBM procedure. But one could argue that these comparisons between GBM and $\hat{\mu}_{OLS}$ are unfair in the following sense: They resemble a situation where the analyst is allowed to fit a rich and flexible π-model but is given no leeway to improve the y-model. We examined many samples of $n = 200$ from this new population and found X_1X_2 to be a strong and highly significant predictor of y in every sample. If we add this one interaction to the y-model, the bias in $\hat{\mu}_{OLS}$ nearly vanishes, and its RMSE becomes comparable to that of the best DR estimators that Ridgeway and McCaffrey tried. Other interactions are often significant as well. We have not examined the performance of $\hat{\mu}_{OLS}$ when these other interactions are included; doing so would be an interesting exercise.

Our point here is not to argue for the superiority of $\hat{\mu}_{OLS}$ over the DR procedures. Either can work well if applied carefully with appropriate safeguards. And either can be made to fail if we, through the design of a simulation, impose artificial restrictions that force the analyst to ignore clear evidence in the observed data that the procedure is flawed.

5. RESPONSE TO ROBINS, SUED, LEI-GOMEZ AND ROTNITZKY

The comments by Robins et al. contain many useful observations and helpful references. Their simulations that reverse the roles of t_i and $1 - t_i$ are instructive. However, in the process of arguing that we misunderstood the message of Bang and Robins (2005), they have apparently misunderstood ours. Their insinuations of cherry-picking might be understandable if we had been arguing for the superiority of $\hat{\mu}_{OLS}$, but that is not what we have done. Quite honestly, we began this investigation fully expecting to demonstrate the *benefits* of dual modeling when neither model is exactly true.

When Bang and Robins (2005) recommended certain DR procedures for routine use, they did so without qualifications or cautionary statements. Now they quote a passage from another article published five years earlier, which Bang and Robins (2005) did not cite, to demonstrate that this was not what they had in mind. Readers cannot react to what they have in mind, but only to what they write. Dr. Robins and his colleagues are eminent researchers, and their statements carry considerable weight. The fact that they knew that these estimators sometimes misbehave but failed to acknowledge it makes their blanket recommendations in 2005 even more troubling.

For the record, we will clarify how we came up with our simulated example. As mentioned in our Section 4, we were trying to loosely mimic a quasi-experiment to assess the average causal effect of dieting on body-mass index among adolescent girls. We decided beforehand that y_i should be predicted from the observed x_i with $R^2 \approx 0.80$, as in the actual data. We decided that the distributions of the estimated propensity scores should resemble those in our Figure 3(e), as in the actual data. We decided that the linear predictors from the y-model and π-model should have a correlation of at least 0.5, as in the actual data, so that $\bar{y}_1 = \sum_i t_i y_i / \sum_i t_i$ would be a strongly biased as an estimator of μ. We decided that the covariates in x_i should not be normally distributed, but they should not be so heavily skewed that a data analyst would need to transform them to reduce the leverage of a few large values. We decided that x_i must be a one-to-one transformation of the unseen true covariates z_i over the effective support of the z_i (without this condition, nonresponse would not be ignorable given x_i). Finally, we decided that the linear regression of y_i and the logistic regression of t_i on x_i would be misspecified to about the same extent, in the sense that the correlations between the linear predictors from each model and the corresponding true linear predictors would be about 0.9.

After considerable trial and error, we came up with one example that met all of these criteria. As we ran our simulations, we were truly surprised to see $\hat{\mu}_{OLS}$ perform as well as it did, consistently beating all competitors. We expected that at least some of the DR estimators would improve upon $\hat{\mu}_{OLS}$, but none did. In fact, we were tempted to look for a different example that would demonstrate some of the benefits of DR, but we decided against it precisely because we wanted to *avoid* cherry-picking.

As Robins et al. deconstruct our simulated example, they suggest that our misspecified linear model $E(y_i) = x_i^T \beta$ is so close to being true that $\hat{\mu}_{OLS}$ is virtually guaranteed to outperform all competitors. If that were so, then why did the DR estimators $\hat{\beta}_{WLS}$ and $\hat{\mu}_{BC\text{-}OLS}$ not perform as well, as those estimators were given the same opportunity to take advantage of this nearly correct y-model? And, if that were so, why would $\hat{\mu}_{OLS}$ perform so poorly in their simulations when the roles of t_i and $1 - t_i$ were reversed?

The first plot in Figure 1 by Robins et al. reveals that (a) the model for y_i given the vector of true covariates z_i is a linear with very high R^2 and (b) the nonresponse is ignorable, so that $P(y_i \mid z_i, t_i = 1)$ and $P(y_i \mid z_i, t_i = 0)$ are the same. This plot implies that conditions where the analyst is allowed to see the z_i's are unrealistic, because knowing z_i is essentially equivalent to knowing y_i. But this plot says nothing about the performance of $\hat{\mu}_{OLS}$ or any other estimator when z_i is hidden and the analyst sees only x_i, which is the only scenario that we have claimed is realistic. [In fact, the first simulated example published by Bang and Robins (2005) yields a similar picture, because their true data-generating mechanism is also linear and their R^2 is 0.94.] The conditional variance $V(y_i \mid z_i)$ was one of many parameters that we had to adjust to create an example that satisfied all of the criteria that we have mentioned. We tried to set $V(y_i \mid z_i)$ to larger values, but doing so decreased the signal-to-noise ratio in the observed data to the point where we no longer saw meaningful biases in any estimators when $n = 200$.

With their Figure 2, Robins et al. purport to show that our misspecified linear regression model fits so well that the predicted values $x_i^T \hat{\beta}$ are essentially unbiased predictions of the missing y_i's, which guarantees excellent performance for $\hat{\mu}_{OLS}$. They state, "We can see that the predicted values of the nonrespondents are reasonably centered around the straight line even for those points with predicted values far from the predicted values of the respondents." On the contrary, our linear model $E(y_i) = x_i^T \beta$ does not give unbiased predictions for nonrespondents or respondents, especially not in the region of extrapolation. To illustrate, we took one simulated sample of $n = 1000$ observations, regressed y_i on x_i among the respondents, and computed the regression predictions $x_i^T \hat{\beta}$ and residuals $y_i - x_i^T \hat{\beta}$ for both groups. A plot of the residuals versus the regression predictions is displayed in Figure 2, along with local polynomial (loess) trends. Respondents are shown in black, and nonrespondents are shown in gray. (For visual clarity, only 20% of the points are displayed, but the loess trends are estimated from the full sample.) For each group, the least-squares regression model strongly underpredicts near the center and overpredicts at the extremes. The reason why $\hat{\mu}_{OLS}$ performs well in this example is not that the linear model is approximately true, but that *the positive and negative residuals in the nonrespondent group approximately cancel out.* The average value of $y_i - x_i^T \hat{\beta}$ for respondents is exactly zero (a consequence of OLS), and the average value of $y_i - x_i^T \hat{\beta}$ for nonrespondents is close to zero. Over 1000 simulated samples, the average of $y_i - x_i^T \hat{\beta}$ among nonrespondents was 1.68. Multiplying this by -0.5 (because the average nonresponse rate is 50%) gives -0.84, the estimated bias for $\hat{\mu}_{OLS}$ reported in our Table 3.

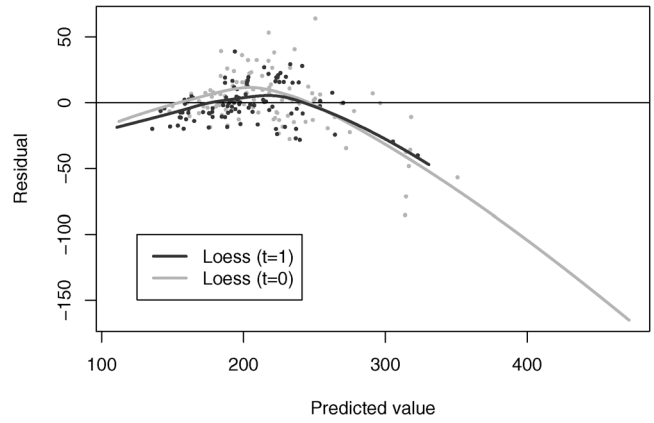


FIG. 2. *Residuals versus predicted values for respondents* ($t_i = 1$) (*black dots*) *and nonrespondents* ($t_i = 0$) (*gray dots*) *from one sample of* $n = 1000$ *from our original simulation, with local polynomial* (*loess*) *trends for each group. For visual clarity, only* 20% *of the sampled points are shown.*

Figure 2 also reveals why $\hat{\mu}_{OLS}$ was not beaten in this example by any of the dual-modeling methods. The differences between the two loess curves in Figure 2 are not large, showing that the OLS predictions have similar patterns of bias for respondents and nonrespondents. When the predictions from a y-model are biased, and the biases are similar when $t_i = 1$ and $t_i = 0$, they are not easily corrected by an estimated propensity model.

If we reverse the roles of t_1 and $1 - t_i$, as Robins et al. have done, the situation dramatically changes. Taking the same sample of $n = 1000$, we regressed y_i on x_i when $t_i = 0$ and predicted the responses for both groups. Residuals versus predicted values from this reverse fit are shown in Figure 3. (Once again, for visual clarity, only 20% of the sampled points are shown, but the loess trends are estimated from the full sample.) For the $t_i = 0$ group, the linear model underpre-

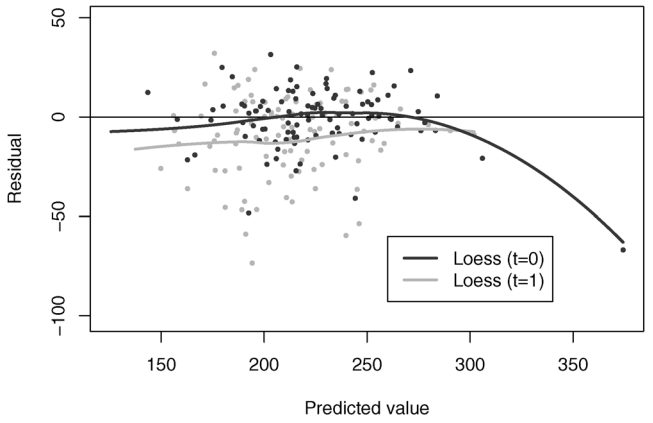


FIG. 3. *Plot analogous to Figure 2, with the roles of t_i and $1 - t_i$ reversed. Cases with $t_i = 0$ and $t_i = 1$ are denoted by black and gray dots, respectively, with local polynomial (loess) trends shown for each group. For visual clarity, only 20% of the sampled points are shown.*

dicts at the center and overpredicts at the extremes, and the average value $y_i = x_i^T \hat{\beta}$ is zero. But for the $t_i = 1$ group, the linear model consistently overpredicts across the entire range, introducing a strong upward bias into $\hat{\mu}_{OLS}$.

This alternative simulation by Robins et al. is a classic example where patterns of bias in a linear y-model cause $\hat{\mu}_{OLS}$ to perform poorly. But because the patterns are dramatically different when $t_i = 0$ and $t_i = 1$, it is also a classic example where the failure can be readily diagnosed and corrected by fitting a π-model. A plot of the residuals $y_i - x_i^T \hat{\beta}$ for the $t_i = 0$ group versus the linear predictors from a logistic propensity model is shown in Figure 4. The plot, which is based only on $(x_i, t_i, (1 - t_i)y_i)$, shows a strong tendency for

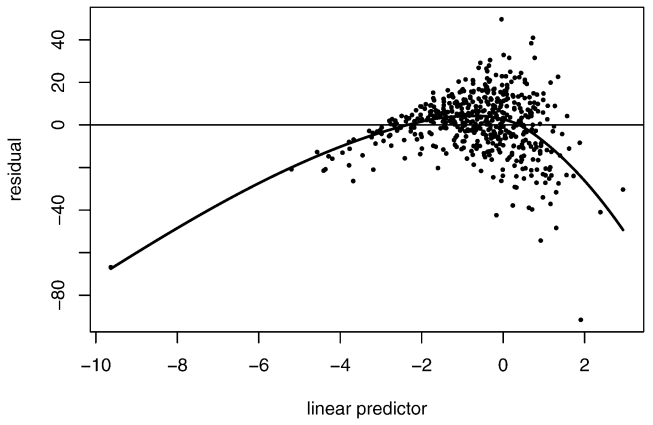


FIG. 4. *Scatterplot of residuals from a linear y-model fit to $t_i = 0$ cases, versus linear predictors from a logistic π-model, with local polynomial (loess) fit in one sample of $n = 1000$ from the alternative simulation study by Robins et al.*

the linear y-model to overpredict when $P(t_i = 1)$ is low or high. To correct this bias, we created a spline basis as in expression (1), with knots at the sample quintiles, and included the four extra terms as predictors in the linear y-model. The performance of $\hat{\mu}_{OLS}$ (which we would now call $\hat{\mu}_{\pi\text{-}cov}$) improved dramatically, and the new estimator worked better than any of the dual-modeling methods reported by Robins et al. The performance statistics in the both-models-wrong scenario were Bias = 2.21, Var. = 12.61 and MSE = 17.46 when $n = 200$, and Bias = 2.40, Var. = 1.88, and MSE = 7.66 when $n = 1000$, which compare favorably to the results shown by Robins et al. in their Table 2.

6. CONCLUDING REMARKS

As statisticians devise newer and fancier methods, we hope to find one that is foolproof, yielding good results no matter when and how it is applied. But the search for a foolproof method is quixotic and futile. Some procedures are, on balance, better than others, but each one requires many subjective inputs, and none should be applied routinely or uncritically. As we develop better estimators, we should also strive to give potential users a healthy dose of intuition about how the procedures work, their limitations, sound recommendations about their use, and diagnostics that can help users decide when a procedure is trustworthy and when it is not.

In conclusion, we believe that propensity modeling is prudent and even necessary when rates of missing information are high. But we are still not convinced that estimated inverse propensities must always be used as weights.

ACKNOWLEDGMENT

This research was supported by National Institute on Drug Abuse Grant P50-DA10075.

REFERENCES

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189

ELLIOTT, M. R. and LITTLE, R. J. A. (2000). Model-based alternatives to trimming survey weights. *J. Official Statistics* **16** 191–209.

KING, G. and ZENG, L. (2002). Improving forecasts of state failure. *World Politics* **53** 623–658.

LITTLE, R. J. A. and AN, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statist. Sinica* **14** 949–968. MR2089342

LUELLEN, J. K., SHADISH, W. R. and CLARK, M. H. (2005). Propensity scores: An introduction and experimental test. *Eval-*

uation Review **29** 530–558.

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

RUBIN, D. B. (2001). Using propensity scores to help design observational studies: Applications to the tobacco litigation. *Health Services and Outcomes Research Methodology* **2** 169–188.

SHAH, B. V., BARNWELL, B. G. and BILER, G. S. (1997). *SUDAAN User's Manual, Release 7.5*. Research Triangle Park, Research Triangle Institute, NC.

TAN, Z. (2006). A distributional approach for causal inference using propensity scores. *J. Amer. Statist. Assoc.* **101** 1619–1637. MR2279484

TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York. MR2233926