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

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).