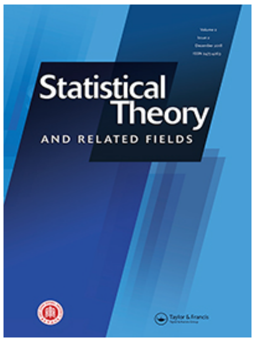

Taylor & Francis
Taylor & Francis Group

Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Semiparametric propensity weighting for nonignorable nonresponse: a discussion of 'Statistical inference for nonignorable missing data problems: a selective review' by Niansheng Tang and Yuanyuan Ju

Jun Shao

STATISTICAL THEORY AND RELATED FIELDS
2018, VOL. 2, NO. 2, 141–142
https://doi.org/10.1080/24754269.2018.1522996

Taylor & Francis
Taylor & Francis Group

SHORT COMMUNICATION

Check for updates

Semiparametric propensity weighting for nonignorable nonresponse: a discussion of 'Statistical inference for nonignorable missing data problems: a selective review' by Niansheng Tang and Yuanyuan Ju

Jun Shao[a,b]

[a]School of Statistics, East China Normal University, Shanghai, People's Republic of China; [b]Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA

ARTICLE HISTORY Received 9 September 2018; Accepted 10 September 2018

Professors Tang and Ju deserve a warm congratulation for their great work on a review of statistical inference for nonignorable missing data problems. Although they called their review 'a selective review', it actually covers most of contemporary advances in the difficult problem of dealing with nonignorable missing data.

In Section 3.1 of Tang and Ju's review, they discussed the weighting approach in estimation with nonignorable nonresponse, which is one of the most popular and effective methods of handling nonresponse. The key in the weighting approach is the estimation of the unknown propensity, the probability of observing the value of a response variable conditional on the value of the response variable and associated covariate values, regardless of whether the response value is observed or not. Tang and Ju reviewed the early developments with parametric models on propensity (e.g., Lee & Tang, 2006; Qin, Leung, & Shao, 2002; Wang, Shao, & Kim, 2014) as well as the more recent advances on semiparametric propensity modelling (Kim & Yu, 2011; Shao & Wang, 2016). The purpose of this note is to add some results and discussions on semiparametric propensity estimation.

We start with some notation. Let y_i be a univariate response variable of interest and x_i be the associated multivariate covariate for the ith sampled unit, $i = 1, \ldots, n$, where y_i is observed if $\delta_i = 1$ and is missing if $\delta_i = 0$, and x_i is always observed. We assume that (y_i, x_i, δ_i), $i = 1, \ldots, n$, are independent and identically distributed. The propensity is defined to be the conditional probability $P(\delta_i = 1|y_i, x_i)$. Since y_i may be missing, this propensity or nonresponse mechanism is nonignorable, and it is ignorable if and only if $P(\delta_i = 1|y_i, x_i) = P(\delta_i = 1|x_i)$.

A parametric model may be imposed on the propensity, but results derived under parametric models may be sensitive to the violations of parametric models, and thus, it is desired to make weaker assumptions. The following semiparametric model is assumed in Kim and Yu (2011),

$$P(\delta_i = 1|y_i, x_i) = \frac{1}{1 + \exp\{g(x_i) + \gamma y_i\}}, \qquad (1)$$

where γ is an unknown parameter and g is an unspecified (nonparametric) function. Note that, under assumption (1), nonresponse is ignorable if and only if $\gamma = 0$, in which case the ignorable propensity $[1 + \exp\{g(x_i)\}]^{-1}$ is nonparametric. Thus, assumption (1) is better than any parametric assumption on propensity because if $\gamma = 0$, any parametric assumption on propensity is unnecessary for handling ignorable missing data.

An extension to (1) is

$$P(\delta_i = 1|y_i, x_i) = \frac{1}{1 + \exp\{g(x_i) + q_\gamma(y_i)\}}, \qquad (2)$$

where q_γ is a parametric function and γ is a possibly multi-dimensional unknown parameter.

As shown in Shao and Wang (2016), under either (1) or (2) the unknown g and γ are not identifiable. Some additional condition is needed to identify the unknown g and γ so that valid estimation and inference is possible. For example, Kim and Yu (2011) assumed that γ is known or can be estimated externally. A more reasonable assumption is to assume that some components of x_i can be excluded from the right-hand side of (2). That is, x_i can be decomposed into u_i and z_i such that

$$P(\delta_i = 1|y_i, x_i) = \frac{1}{1 + \exp\{g(u_i) + q_\gamma(y_i)\}}, \qquad (3)$$

while z_i, the part of the covariate vector not in the right-hand side of (3), is still a useful covariate in the sense that the conditional distribution of y_i given x_i depends on z_i. This idea was developed in Wang et al. (2014), Zhao and Shao (2015), and Shao and Wang (2016), where they named the covariate z_i to be a nonresponse instrument.

CONTACT Jun Shao shao@stat.wisc.edu

Following Tang and Ju's review and Shao and Wang (2016), we can show that assumption (3) implies that

$$\exp\{g(u_i)\} = \frac{E(1-\delta_i|u_i)}{E[\delta_i \exp\{q_\gamma(y_i)\}|u_i]} \quad (4)$$

and

$$E\left[h(z_i)\left\{\delta_i[1+\exp\{g(u_i)+q_\gamma(y_i)\}]-1\right\}\right]=0, \quad (5)$$

where $h(z_i)$ is a vector function of z_i. Under suitable conditions on $h(z_i)$, asymptotically valid estimators of g and γ can be obtained based on (4) and (5), using either the method of generalised moments (Shao & Wang, 2016) or the empirical likelihood method in Tang and Ju's review (Section 3.2).

Once g and γ are estimated, the weighting approach using the inverse of propensity in (3) with g and γ replaced by their estimators can be applied to estimate parameters of interest in the distribution of y_i or the conditional distribution of y_i givenx_i.

Now, consider changing assumption (3) to

$$P(\delta_i = 1|y_i, x_i) = g(u_i)q_\gamma(y_i), \quad (6)$$

where g is nonparametric, q_γ is parametric, and both are between 0 and 1. Note that (6) is a multiplicity model considered by Zhao and Shao (2017). Under (6),counterparts of (4) and (5) are, respectively,

$$g(u_i) = E\left\{\frac{\delta_i}{q_\gamma(y_i)}\middle| u_i\right\} \quad (7)$$

and

$$E\left[h(z_i)\left\{\frac{\delta_i}{g(u_i)q_\gamma(y_i)}-1\right\}\right]=0. \quad (8)$$

Asymptotically valid estimators of g and γ can be obtained using (7) and (8) and similar techniques in Shao and Wang (2016).

Alternatively, we may change (3) to

$$P(\delta_i = 1|y_i, x_i) = \frac{1}{1+g(u_i)+q_\gamma(y_i)}, \quad (9)$$

where g is nonparametric, q_γ is parametric, and $g(u_i) + q_\gamma(y_i)$ is between 0 and 1. Note that the difference between model (3) and model (9) is that the former has a multiplicity effect of $g(u_i)$ and $q_\gamma(y_i)$ on propensity, whereas the latter has an additive effect of $g(u_i)$ and $q_\gamma(y_i)$. Under (9), counter parts of (4) and (5) are, respectively,

$$g(u_i) = \frac{1-E[\delta_i\{1+q_\gamma(y_i)\}|u_i]}{E(\delta_i|u_i)} \quad (10)$$

and

$$E\left[h(z_i)\left\{\delta_i[1+g(u_i)+q_\gamma(y_i)]-1\right\}\right]=0. \quad (11)$$

Again, asymptotically valid estimators of g and γ can be obtained using (10) and (11) and similar techniques in Shao and Wang (2016).

We conclude with the following question. What is a general assumption for which (3), (6) and (9) are all special cases and results similar to (4) and (5) can be derived?

Consider

$$P(\delta_i = 1|y_i, x_i) = \pi\{g(u_i), q_\gamma(y_i)\}, \quad (12)$$

where g is nonparametric, q_γ is parametric, and $\pi\{\cdot,\cdot\}$ is a two-dimensional known function. Note that (3), (6) and (9) are all special cases of (12). The previous results (4), (7) and (10) are all derived based on

$$E\left[\frac{\delta_i}{\pi\{g(u_i), q_\gamma(y_i)\}}\middle| u_i\right] = E\left[E\left\{\frac{\delta_i}{\pi\{g(u_i), q_\gamma(y_i)\}}\middle| y_i, u_i\right\}\middle| u_i\right] = 1.$$

Define

$$\psi\{g(u_i), u_i\} = E\left[\frac{\delta_i}{\pi\{g(u_i), q_\gamma(y_i)\}}\middle| u_i\right], \quad (13)$$

a function of $g(u_i)$ and u_i. If, for each fixed u_i, ψ is a strictly monotone function of $g(u_i)$, then (13) defines a possibly implicit function of $g(u_i)$ and similar results may be derived.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was supported by the Chinese 111 Project, the Fundamental Research Funds for the Central Universities and the U.S. National Science Foundation (Directorate for Mathematical and Physical Sciences) grant DMS-1612873.

References

Kim, J. K., & Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association*, *106*, 157–165.

Lee, S. Y., & Tang, N. S. (2006). Bayesian analysis of nonlinear structural equation models with nonignorable missing data. *Psychometrika*, *71*, 541–564.

Qin, J., Leung, D., & Shao, J. (2002). Estimation with survey data under non-ignorable nonresponse or informative sampling. *Journal of American Statistical Association*, *97*, 93–200.

Shao, J., & Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, *103*, 175–187.

Wang, S., Shao, J., & Kim, J. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, *24*, 1097–1116.

Zhao, J., & Shao, J. (2015). Semiparametric pseudo likelihoods in generalized linear models with nonignorable missing data. *Journal of American Statistical Association*, *110*, 1577–1590.

Zhao, J., & Shao, J. (2017). Approximate conditional likelihood for generalized linear models with general missing data mechanism. *Journal of System Science and Complexity*, *30*, 139–153.