The identification problem for nonignorable nonresponse data

Kosuke Morikawa, Osaka University
Jae Kwang Kim, Iowa State University

When the response mechanism is believed to be not missing at random (NMAR), a valid analysis requires stronger assumptions on the response mechanism than standard statistical methods. Let Y be an outcome variable subject to missingness, X be a covariate vector, and R be the response indicator of Y, which is 1(0) when Y is observed (missing). In this setup, the NMAR mechanism is characterized by conditional dependence: R depends on Y even after controlled for X. For the model identification under NMAR, both outcome $f(y \mid x; \beta)$ and response mechanism $\pi(x, y; \phi) := P(R = 1 \mid x, y; \phi)$ must be restricted, where $f(y \mid x; \beta)$ is conditional distribution of y given x known up to a finite or infinite dimensional parameter β, and $\pi(x, y; \phi)$ is a response model known up to a finite dimensional parameter ϕ.

We wish to show that $\pi(x, y; \phi)f(y \mid x; \beta) = \pi(x, y; \phi')f(y \mid x; \beta')$ for almost all (x, y) implies $\phi = \phi'$ and $\beta = \beta'$. To guarantee this property, Wang et al. (2014) assumed that there exists a nonresponse instrumental variable x_1 in $x = (x_1^{\mathrm{T}}, x_2^{\mathrm{T}})^{\mathrm{T}}$such that x_2 is independent of r, given x_1 and y. A graphical model of this dependence is shown in Figure 1. Although the existence of such a nonresponse instrumental variable is a sufficient condition, it is hard to prove with observed data. In the meantime, Miao et al. (2016) derived a sufficient condition for model identification by putting restriction on the outcome model being normal or normal mixture model. Although their condition does not require the instrumental variable, it is hard to verify the outcome model being normal or normal mixture from observed data.

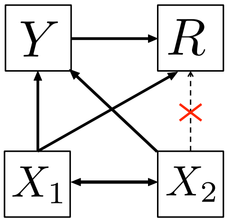

Figure 1: Nonresponse instrumental variable.

Therefore, we propose a necessary and sufficient condition for the model identification by restricting the distribution of $f_1(y \mid x; \gamma) := f(y \mid x, r = 1; \gamma)$, where γ is a finite or infinite dimensional parameter (Morikawa & Kim , 2017). Unlike Miao et al. (2016)'s condition, our condition is necessary and sufficient and can be checked with observed data because the restriction is on the observed distribution $f_1(y \mid x)$. Furthermore, it does not require the use of instrumental variable. For example, suppose that $f_1(y \mid x)$ is normal distribution with mean $\tau(x)$ and variance σ^2, and the response model is $\mathrm{logit}\{\pi(x, y; \phi)\} = \phi_0 + \phi_1 x + \phi_2 y$. By using our condition, we can show that this model is identifiable unless $\tau(x)$ is linear. We will present an application to Korean Labor and Income Panel Survey data.

References

Miao, W., Ding, P. & Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *J. Am. Statist. Assoc.* **111**, 1673–1683.

Morikawa, K. & Kim, J. K. (2017). Semiparametric Adaptive Estimation With Nonignorable Nonresponse Data. Submitted.

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