Semiparametric Adaptive Estimation With Nonignorable Nonresponse

*Kosuke Morikawa[1], Jae-kwang Kim[2]

1. Earthquake Research Institute, The University of Tokyo, 2. Iowa State University, Department of Statistics

Statistical inference with missing data has become a major issue in many empirical research fields including medical research, econometrics, psychometrics and seismology. If data involve missing values, conventional statistical methods cannot be directly applied. In this talk, we study nonresponse, a typical type of missingness, in which a response variable (or outcome variable) is subject to missingness.

A key concept for valid analysis with missing data is response mechanism or missing-data mechanism, that is also called propensity score especially in epidemiology. This mechanism is mathematically defined as a conditional probability of the response variable being observed given all other variables. It is important to specify the response mechanism in missing data analysis. Therefore, the response mechanism is needed to be modeled by a parametric model, which is called response model, and estimated.

When missingness depends on the missing value, the mechanism is said to be nonignorable; most missing data are nonignorable nonresponses and this type of missingness is most difficult to handle. Appropriate analysis of nonignorable nonresponse data requires strong unverified assumptions such as correct specification of outcome model of complete data or existence of instrumental variables. It is hard to specify a response model in general; even though it can be specified, identifiability of the response model often fails and indeed it is even difficult to check the identifiability.

The first contribution of this talk is to introduce a semiparametric approach to estimate a response model to overcome the difficulties described above. We propose two types of semiparametric estimators. The first estimator requires correct specification of an outcome model of observed data in addition to the response model. Modeling the outcome model is more reasonable than classical methods because it is a model for the observed data we have. Furthermore, the estimator has consistency and asymptotic normality regardless of whether the specification is correct or wrong. The second estimator does not require any model other than the response model. Both two estimators can attain the semiparametric efficiency bound, which is the information bound only when the response mechanism is modeled.

The second contribution is to provide useful conditions for checking the model identifiability in the analysis of nonignorable nonresponse data. The condition can be checked by using observed data only, and do not rely on any instrumental variables. Based on the conditions, some identifiable models are proposed to analyze nonignorable nonresponse data.

Numerical experiments are conducted to show that our semiparametric estimators outperform other existing estimators in terms of bias and variance.

Keywords: Identification, Incomplete data, Not missing at random (NMAR), Semiparametric efficient estimation

