Statistica Sinica **36** (2026), 1-21
doi:https://doi.org/10.5705/ss.202023.0383

ON DOUBLY ROBUST ESTIMATION WITH NONIGNORABLE MISSING DATA USING INSTRUMENTAL VARIABLES

Baoluo Sun*, Wang Miao and Deshanee S. Wickramarachchi

National University of Singapore, Peking University and University of Colombo

Abstract: Suppose we are interested in the mean of an outcome that is subject to nonignorable nonresponse. This paper develops new semiparametric estimation methods with instrumental variables which affect nonresponse, but not the outcome. The proposed estimators remain consistent and asymptotically normal even under partial model misspecifications for two variation independent nuisance components. We evaluate the performance of the proposed estimators via a simulation study, and apply them in adjusting for missing data induced by HIV testing refusal in the evaluation of HIV seroprevalence in Mochudi, Botswana, using interviewer experience as an instrumental variable.

Key words and phrases: Doubly robust estimation, endogeneous selection, exclusion restriction, instrumental variable, nonignorable nonresponse.

1. Introduction

Missing data are ubiquitous in the health and social sciences. Our motivating example concerns a household survey in Mochudi, Botswana to estimate HIV seroprevalence among adults. About 19% of the adults who were contacted for the survey have missing final HIV status, mainly due to refusal to participate in the HIV testing component. The nonresponse is said to be ignorable if, conditional on the fully observed variables, it is independent of the underlying HIV status (Rubin, 1976; Rubin and Little, 2019). In this case, the HIV seroprevalence among respondents is representative of the overall HIV seroprevalence in the population, within strata of the observed variables. Nonetheless, ignorability is a strong assumption which may be untenable in practice; for instance, HIV testing refusal may be entangled with features of the underlying HIV status in the household survey. The problem of nonignorable nonresponse has therefore received much attention in the missing data literature. Robins, Rotnitzky and Scharfstein (2000) described a general class of models which requires *a priori* specification of a selection bias parameter that encodes the residual association of the nonresponse mechanism with the outcome of interest, after adjusting for the fully observed variables. It coincides with the widely adopted exponential

*Corresponding author. E-mail: stasb@nus.edu.sg