Statistica Sinica: Preprint
doi:10.5705/ss.202016.0324

Missing data occur frequently in empirical studies in health and social sciences, often compromising our ability to make accurate inferences. An outcome is said to be missing not at random (MNAR) if, conditional on the observed variables, the missing data mechanism still depends on the unobserved outcome. In such settings, identification is generally not possible without imposing additional assumptions. Identification is sometimes possible, however, if an instrumental variable (IV) is observed for all subjects which satisfies the exclusion restriction that the IV affects the missingness process without directly influencing the outcome. In this paper, we provide necessary and sufficient conditions for nonparametric identification of the full data distribution under MNAR with the aid of an IV. In addition, we give sufficient identification conditions that are more straightforward to verify in practice. For inference, we focus on estimation of a population outcome mean, for which we develop a suite of semiparametric estimators that extend methods previously developed for data missing at random. Specifically, we propose inverse probability weighted estimation, outcome regression-based estimation and doubly robust estimation of the mean of an outcome subject to MNAR. For illustration, the methods are used to account for selection bias induced by HIV testing refusal in the evaluation of HIV seroprevalence in Mochudi, Botswana, using interviewer characteristics such as gender, age and years of experience as IVs.