Imputation methods for a binary variable

Seppo Laaksonen
University of Helsinki: e-mail: Seppo.Laaksonen@Helsinki.Fi

Abstract

Binary variables are common in surveys including such as employed vs unemployed, healthy vs unhealthy or poor vs non-poor. The last one is used in the examples of this paper. It is unfortunate that survey data are never complete, that is, missingnessoccurs , sometimes severely. In those cases missingness may violate estimates due to unit nonresponse. For this purpose, there are also sophisticated nonresponse adjustments possible to use, and these should be applied in a best way taking advantage of auxiliary variables from the nonrespondents as well as from the population level aggregates. In the case when the nonrespondents participate partially in the survey (item nonresponse), there exists more micro level auxiliary variables available. Such data should naturally be exploited in a best way as well. This leads to apply imputation methods. In this paper, our binary variable gives a too low poverty rate when estimating it from the completely responded respondents. So it is beneficial to improve the estimate with imputations. This is not necessarily easy since our pattern of auxiliary data is not excellent that is common in real life. Consequently, the results depend more than in a nice situation on an imputation method applied. In this case we compare several methods. The imputation process consists of the two steps, (i) imputation model and (ii) imputation task. Interestingly, the dependent variable of the imputation model is also binary but they are of two types: (a) the variable being imputed itself, or (b) the binary response indicator. The former takes advantage of the data of the respondents but the latter both from the respondents and from the nonrespondents. Since we know the true values (but so that their mechanism is not known), we can easily compare different methods. When using the methods with random numbers, we also can apply multiple imputation methodology. So we have both single imputation and multiple imputation methods compared in the empirical part. Our strategy is not only Bayesian for multiple imputation that is usual in software packages such as SPSS and SAS. We test their methods but concentrate on our own solutions, called non-Bayesian. The key words include some specific methodologies applied.

Keywords:Model-donor imputation methods, Real-donor imputation methods, logistic regression, probit regression