Journal of the Korean Statistical Society 41 (2012) 291–303

Contents lists available at SciVerse ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Parametric fractional imputation for nonignorable missing data

Ji Young Kim [a], Jae Kwang Kim [b,*]

[a] Department of Applied Statistics, Yonsei University, Seoul, 120-749, Republic of Korea

[b] Department of Statistics, Iowa State University, Ames, IA 50011, USA

ARTICLE INFO

Article history:
Received 21 April 2011
Accepted 18 October 2011
Available online 9 November 2011

AMS 2000 subject classifications:
62F99
62D99

Keywords:
EM algorithm
Monte Carlo EM
Multiple imputation
Not missing at random

ABSTRACT

Parameter estimation with missing data is a frequently encountered problem in statistics. Imputation is often used to facilitate the parameter estimation by simply applying the complete-sample estimators to the imputed dataset.

In this article, we consider the problem of parameter estimation with nonignorable missing data using the approach of parametric fractional imputation proposed by Kim (2011). Using the fractional weights, the E-step of the EM algorithm can be approximated by the weighted mean of the imputed data likelihood where the fractional weights are computed from the current value of the parameter estimates. Calibration fractional imputation is also considered as a way for improving the Monte Carlo approximation in the fractional imputation. Variance estimation is also discussed. Results from two simulation studies are presented to compare the proposed method with the existing methods. A real data example from the Korea Labor and Income Panel Survey (KLIPS) is also presented.

1. Introduction

Missing data is frequently encountered in many areas of statistics. Statistical analysis in the presence of missing data has been an area of considerable interest because simply ignoring the missing part of the data often destroys the representativeness of the remaining sample. If the missing probability is unrelated to the unobserved values after adjusting for the effect of auxiliary variables in the model, the nonresponse is called missing at random (MAR) or ignorable. Nonignorable, or non-MAR model refers to the situation where missingness depends directly on the missing values. For example, in the survey of income, some individuals are reluctant to answer because they have high income. If the probability of not observing the income amount is associated with the value of income, even after controlling for observed covariates such as age and occupation, then the response mechanism is not ignorable and it has to be explicitly modeled for valid inference. Little and Rubin (2002) and Molenberghs and Kenward (2007) provide a comprehensive overview of the missing data analysis.

Parameter estimation under nonignorable missing is a challenging problem because the response mechanism is generally unknown. Nordheim (1984) showed that if some information of the probabilities of uncertain classification is obtained, then the category is identified under the nonignorable missing data mechanism. Baker and Laird (1988) used the EM algorithm to estimate the maximum likelihood estimators of the expected cell counts under a log-linear model for categorical missing data with nonignorable missing. Park and Brown (1997) proposed the maximum likelihood estimating method with constraints for categorical data using a data-dependent prior, which amounts to adding additional observation for the missing data.

Imputation is a process of assigning values for the missing responses to produce a complete data set. Reasons for conducting imputation are to facilitate analyses using complete data analysis methods, to ensure that the results obtained

* Corresponding author.
E-mail address: jkim@iastate.edu (J.K. Kim).

1226-3192/$ – see front matter
doi:10.1016/j.jkss.2011.10.002

by different analyses are consistent with one another, and to reduce nonresponse bias. Imputation is a technique of handling missing data for general purpose estimation in the sense that the imputed data can be used to estimate several parameters (Meng, 1994). Thus, imputation is very popular in handling missing data for public-use data because it provides consistent results among different users and reduces the burden of modeling missing data for the ultimate data users (Rubin, 1996). Imputation for general purpose estimation under nonignorable nonresponse is rarely addressed in the literature. Greenlees, Reece, and Zieschang (1982) proposed an iterative method of imputation but they did not address general purpose estimation. Glynn, Laird, and Rubin (1993) discussed the application of multiple imputation to nonignorable missing data using the so-called pattern mixture model of Little (1993).

In this article, we proposed a new imputation method for general purpose estimation based on the parametric fractional imputation (PFI) method of Kim (2011). In Kim (2011), the PFI method was developed under ignorable nonresponse. Thus, a detailed discussion of the PFI method to nonignorable missing is desired. One advantage of the PFI method is that, if the imputed data is applied to the score function, the resulting estimator is very close to the maximum likelihood estimator. Thus, the PFI method leads to more efficient estimates than multiple imputation or the iterative method of Greenlees et al. (1982).

The paper is organized as follows. In Section 2, basic setup is introduced. In Section 3, we propose a parametric estimation method under the nonignorable missing data mechanism using the parametric fractional imputation. In Section 4, some computational details are discussed. In Section 5, the variance estimation method is proposed. In Section 6, results from two simulation studies are presented. A real data example using the Korea Labor and Income Panel Survey (KLIPS) is presented in Section 7.

2. Basic setup

For simplicity, consider two variables, $\mathbf{x}$ and y, where $\mathbf{x}$ is always observed and y is subject to missing. Assume that we have a parametric model for the conditional distribution of y given $\mathbf{x}$, denoted by $f_1(y \mid \mathbf{x}, \theta)$ with unknown parameter θ, and the marginal distribution of $\mathbf{x}$ is completely unspecified. The parameter of interest is the parameter θ in the conditional distribution

$$y_i \mid \mathbf{x}_i \sim f_1(y \mid \mathbf{x}, \theta). \tag{1}$$

Under complete response, the maximum likelihood estimator of θ can be obtained by solving the following score equation for θ:

$$S_1(\theta) \equiv n^{-1} \sum_{i=1}^{n} S_1(\theta; \mathbf{x}_i, y_i) = 0$$

where $S_1(\theta; \mathbf{x}, y) = \partial \log f_1(y \mid \mathbf{x}, \theta)/\partial\theta$.

Under the existence of nonresponse, let r_i be the response indicator for y_i, defined by

$$r_i = \begin{cases} 1 & y_i \text{ is observed} \\ 0 & y_i \text{ is missing.} \end{cases}$$

We assume that the response mechanism is independent and

$$r_i \mid (\mathbf{x}_i, y_i) \sim \textit{Bernoulli}\,(\pi_i) \tag{2}$$

where $\pi_i = \pi(\mathbf{x}_i, y_i; \phi)$ for some function $\pi(\cdot)$ known up to ϕ. Thus, the conditional density of r_i given $\mathbf{x}_i$ and y_i is

$$f_2(r \mid \mathbf{x}, y; \phi) = \{\pi(\mathbf{x}, y; \phi)\}^r \{1 - \pi(\mathbf{x}, y; \phi)\}^{1-r}.$$

If y_i were also available, then the maximum likelihood estimator of ϕ can be obtained by solving the following score equation for θ:

$$S_2(\phi) \equiv n^{-1} \sum_{i=1}^{n} \{r_i - \pi(\mathbf{x}_i, y_i; \phi)\} \partial \text{logit}\{\pi(\mathbf{x}_i, y_i; \phi)\}/\partial\phi = 0$$

where $\text{logit}(p) = \log\{p/(1-p)\}$. Thus, the joint score function for (θ, ϕ) is

$$S_n(\theta, \phi)' = [S_1(\theta)', S_2(\phi)']. \tag{3}$$

When we observe $(\mathbf{x}_i, r_i y_i, r_i)$ in the sample, the observed likelihood of (θ, ϕ) is

$$L_{obs}(\theta, \phi) = \prod_{i=1}^{n} f_{obs}(r_i y_i, r_i \mid \mathbf{x}_i; \theta, \phi),$$

where

$$f_{obs}(r_i y_i, r_i \mid \mathbf{x}_i; \theta, \phi) = \begin{cases} f_1(y_i \mid \mathbf{x}_i; \theta) f_2(r_i \mid \mathbf{x}_i, y_i; \phi) & \text{if } r_i = 1 \\ g(\mathbf{x}_i, r_i; \theta, \phi) & \text{if } r_i = 0 \end{cases}$$