Statistics and Probability Letters 107 (2015) 204–209

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

A new multiple imputation method for bounded missing values

Tae Yeon Kwon [a], Yousung Park [b,*]

[a] Institute for Economic Research, Korea University, Republic of Korea
[b] Department of Statistics, Korea University, Republic of Korea

ARTICLE INFO

Article history:
Received 10 November 2014
Received in revised form 28 August 2015
Accepted 28 August 2015
Available online 7 September 2015

Keywords:
Multiple imputation
Boundary condition
Hot-deck
Proportioned residuals
Boundary information matching

ABSTRACT

A proportioned residual draw method is introduced to impute bounded missing values. It is shown that this new regression-based imputation method is least biased and robust for heavy-tailed and skewed error distributions, regardless of number of boundaries and missing mechanisms.

1. Introduction

Multiple imputation (Rubin, 1978) substitutes each missing value with a set of plausible values that are obtained from observed data, resulting in multiple completed data sets that allow imputation uncertainty to be incorporated into statistical inferences. Regression-based multiple imputation which we focus on in this paper replaces each missing regression outcome with a set of values randomly drawn from a predictive or an empirical distribution.

The normal imputation method (NM) uses a posterior normal distribution based on the regression coefficients to impute missing data (Rubin, 1978), whereas the method adjusted for uncertainty of mean and variance (MV) uses the empirical distribution of the standardized residual in such a way that each missing value is imputed by its predictive mean plus the residual that corresponds to the randomly chosen standardized residual (Rubin and Schenker, 1986; Rubin, 2004).

Partially parametric imputation methods have their origins in the "hot-deck" method in which a missing value is imputed from nearby observed cases. The predictive mean matching method (PMM) imputes a missing value from an observation randomly drawn from a set of observed cases (or, equivalently, possible donors) having predictive means close to the predictive mean of the missing value (Schenker and Taylor, 1996). On the other hand, local residual draw methods (LRD) impute each missing value using its predictive mean plus a residual that is randomly drawn from the residuals of a set of observed cases with predictive means close to that of the missing value (Schenker and Taylor, 1996). Three matching types used in PMM and LRD are compared in Morris et al. (2014).

We propose a new residual drawing technique for regression-based multiple-imputation when missing values are bounded. Typical survey data consist of a large number of variables falling into certain logical or consistency bounds imposed by survey questionnaires, as mentioned in He and Raghunathan (2009). As advanced ceramics (AC) is emerging industry,

* Corresponding author.
E-mail address: yspark@korea.ac.kr (Y. Park).

http://dx.doi.org/10.1016/j.spl.2015.08.028
0167-7152/

statistics has been not yet built up, but AC manufacturing companies are mandated to report their total sales to a public institution such as financial supervisory service in Korea. Thus the AC sales, which are missing but to be estimated, are bounded by corresponding total sales. Smoking periods of smokers are often missing in health screening questionnaire and the missing smoking period of each smoker is bounded by his/her age.

A typical way to reflect such a boundary condition in the imputation procedure is to include an acceptance/rejection step. However, this additional step leads to a biased estimate in the existing imputation methods. Our new imputation method guarantees such a boundary condition by using the residual divided by the distance between the predictive mean and its boundary value, called a proportioned residual. Partially parametric imputation methods using proportioned residuals are also proposed to improve the performance of our new imputation method for a distribution that is highly skewed and/or heavy-tailed due to boundary conditions.

This paper consists of five sections. We describe our new imputation approach in Section 2 and show how this approach ensures that the boundary conditions hold without an additional acceptance/rejection step. In Section 3, we perform extensive simulation studies to compare our imputation method to NM, MV, PMM, and LRD for three different error distributions, showing that our method produces the least-biased predictive means and is robust for heavy-tailed and skewed error distributions. We apply our method to real data analysis in Section 4, which is the motivation for this paper. Section 5 includes concluding remarks.

2. Proportioned residual draw method (PRD)

Regression-based multiple imputation starts with the generation of the regression coefficients β^* and variance σ^{*2} from the posterior distributions given by

$$\sigma^{*2} \sim \hat{\sigma}^2_{OLS}(n_{obs} - q)/\chi^2_{n_{obs}-1}, \qquad \beta^* \sim N\left(\hat{\beta}_{OLS}, \sigma^{*2}(X^TX)^{-1}\right) \tag{1}$$

where X is the fully observed q covariates, and $\hat{\beta}_{OLS}$ and $\hat{\sigma}^2_{OLS}$ are the OLS estimates of regression coefficients and variance, respectively, from the regression model: $Y^{obs} = X^T\beta + \varepsilon$, where Y^{obs} is observed values of Y with size n_0. A comprehensive development of (1) is given in Rubin (1987).

Our multiple imputation method also follows this step to draw β^* and σ^*. Let C_i be the known upper bound of Y_i, $i = 1, 2, \ldots, n_0, n_0 + 1, \ldots, n$, and hence the first n_0 Y_is are observed and the remaining $(n - n_0)$ Y_is are missing. Define the proportioned residual as given by

$$\tilde{r}_i = \frac{Y_i - \hat{Y}_i^{obs}}{C_i - \hat{Y}_i^{obs}}, \quad i = 1, \ldots, n_0 \text{ and } Y_i - \hat{Y}_i^{obs} \le C_i - \hat{Y}_i^{obs}, \tag{2}$$

where $\hat{Y}_i^{obs} = X_i^T\beta^{obs}$ and $C_i - \hat{Y}_i^{obs} \neq 0$. According to the sign of $C_i - \hat{Y}_i^{obs}$, we divide the proportioned residuals given in (2) into two sets:

$$\tilde{R}^+ = \{\tilde{r}_i \text{ with } C_i - \hat{Y}_i^{obs} > 0\} \quad \text{and} \quad \tilde{R}^- = \{\tilde{r}_i \text{ with } C_i - \hat{Y}_i^{obs} < 0\}. \tag{3}$$

Then the proportioned residual draw (PRD) method imputes the $(n - n_0)$ missing Y_js with Y_j^*s as follows.

$$Y_j^* = X_j^T\beta^{mis} + \tilde{r}_j^*(C_j - \hat{Y}_j^{mis}) \quad \text{for } j = n_0 + 1, \ldots, n, \tag{4}$$

where $\hat{Y}_j^{mis} = X_j^T\beta^{mis}$ and $\tilde{r}_j^*$ is randomly chosen from $\tilde{R}^+$ if $C_j - \hat{Y}_j^{mis} \ge 0$ and $\tilde{R}^-$ if $C_j - \hat{Y}_j^{mis} < 0$ for $j = n_0 + 1, \ldots, n$.

Summarizing the above, the PRD procedure consists of the following four steps:

Step 1. Draw σ^{*2} and β^* from the respective distribution given in (1) and let $\beta^{obs} = \beta^{mis} = \beta^*$.
Step 2. Define $\tilde{r}_i$ as possible donors and sets $\tilde{R}^+$ and $\tilde{R}^-$ as given by (2) and (3)
Step 3. Randomly select $\tilde{r}_j^*$ from $\tilde{R}^+$ or $\tilde{R}^-$ according to the sign of $C_j - \hat{Y}_j^{mis}$.
Step 4. Impute the missing Y_j with Y_j^* calculated by (4).

These four steps are repeated M times for multiple imputation.

Theorem 2.1. *The imputed value Y_j^* in Step 4 satisfies its boundary condition.*

(1) *Y_j^* in Step 4 is less than C_j, for $j = n_0 + 1, \ldots, n$.*
(2) *If the Y s are lower bounded, rather than upper bounded, the PRD described above also ensures that Y_j^* in Step 4 is greater than C_j.*

The proof is given in the Appendix. When we let $\beta^{obs} = \beta^{mis} = \hat{\beta}_{OLS}$ or $\beta^{obs} = \hat{\beta}^{OLS}$ and $\beta^{mis} = \beta^*$, the matching is referred to as type 0 or type 1 matching, respectively (Morris et al., 2014). Our case (*i.e.*, $\beta^{obs} = \beta^{mis} = \beta^*$) is called the type 2 matching. Since, in proving Theorem 2.1, $\tilde{r}_j^*$ in Step 3 is most important and free from the matching type, Theorem 2.1 holds for all the three types of matching.