Journal of Statistical Planning and Inference 176 (2016) 33–51

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Multiple imputation in three or more stages

J. McGinniss [a,*,1], O. Harel [b]

[a] Boehringer Ingelheim Pharmaceuticals, Ridgefield, CT, USA

[b] Department of Statistics, University of Connecticut, Storrs, CT, USA

ARTICLE INFO

Article history:
Received 1 October 2015
Received in revised form 5 April 2016
Accepted 7 April 2016
Available online 20 April 2016

Keywords:
Multiple imputation
Missing data
Ignorability

ABSTRACT

Missing values present challenges in the analysis of data across many areas of research. Handling incomplete data incorrectly can lead to bias, over-confident intervals, and inaccurate inferences. One principled method of handling incomplete data is multiple imputation. This article considers incomplete data in which values are missing for three or more qualitatively different reasons and applies a modified multiple imputation framework in the analysis of that data. Included are a proof of the methodology used for three-stage multiple imputation with its limiting distribution, an extension to more than three types of missing values, an extension to the ignorability assumption with proof, and simulations demonstrating that the estimator is unbiased and efficient under the ignorability assumption.

1. Introduction

Incomplete data are a common obstacle to the analysis of data in a variety of fields, ranging from clinical trials to social sciences (Molenberghs and Kenward, 2007). Missing values can occur for several different reasons including failure to answer a survey question, dropout, planned missing values, intermittent missed measurements, latent variables, and equipment malfunction. In fact, many studies will have more than just one type of missing value. Appropriately handling missing values is critical in the inference for a parameter of interest (Belin, 2009; DeSouza et al., 2009). Many methods of handling missing values inappropriately fail to account for the uncertainty due to missing values (Wood et al., 2004; Harel et al., 2012). This failure to account for uncertainty can lead to biased estimates and over-confident inferences.

Multiple imputation is one method for handling incomplete data that accounts for the variability of the incomplete data. This procedure does so by filling in plausible values several times to create several complete data sets and then appropriately combining complete data estimates using specific combining rules (Rubin, 1987). This method is praised for the ability to use complete data analytical methods on each data set as well as for retaining the variability seen in the observed data to arrive at estimates which are not distorted by the imputation method (Schafer, 1997; Schafer and Graham, 2002; Harel and Zhou, 2007; White and Carlin, 2010).

In practice, there is often more than one type of missing value in a study. Knowledge about the nature of the missing values can help identify the most appropriate method for dealing with missing data (Little and Rubin, 2002). Typically, all of the missing values are treated as though they are the same type. However, there are benefits to treating each of these types of missing values separately. One benefit arises when imputing one type of missing value first computationally simplifies the imputation of the rest of the missing values (Shen, 2000). A second important benefit is that treating the two types

* Corresponding author. Tel.: +1 203 791 6476.
E-mail addresses: jennifer.mcginniss@boehringer-ingelheim.com (J. McGinniss), ofer.harel@uconn.edu (O. Harel).

[1] née Boyko.

http://dx.doi.org/10.1016/j.jspi.2016.04.001
0378-3758/

differently can allow the researcher to quantify how much variability and how much missing information is due to each type of missing value (Harel, 2007). A third benefit is the ability to assign different missing data assumptions to each of the missing value types (Harel and Schafer, 2009). That is, each of the missing value types might have different assumptions on the mechanisms generating the missingness. Two-stage multiple imputation is a nested version of multiple imputation where missing values of one type are imputed first (Rubin, 2003; Kinney and Reiter, 2009). For each imputation of the first type of missing value, additional imputations are created for the second type of missing value, holding the first imputed values fixed. Separate combining rules determine how estimates should be handled to lead to valid inferences (Shen, 2000).

One area which is still unexplored is the situation where there are three or more types of missing values in a study. This is a natural extension of two-stage multiple imputation which allows for the flexibility required when dealing with real-world data. Studies which tend to have large amounts of missing values also tend to have missing values of several different types. One practical example is that of a longitudinal clinical trial where there may be missing values due to a patient being lost to follow-up, a patient missing visits intermittently over the course of a trial, and corrupted (and therefore, missing) laboratory measurements. A second example which is more relevant to surveys would be missing values due to item nonresponse, unit nonresponse, and a missing latent class. Development of a three-stage multiple imputation approach is beneficial in analyzing both of these types of studies. Three-stage multiple imputation also extends the benefits of two-stage multiple imputation, namely the quantification of the variability attributable to each type of missing value and the flexibility for greater specificity regarding data analysis. This work seeks to better simulate the reality of data collection in practice. Missing data are rarely straightforward enough to be of one, or even two, types. Multiple imputation in three stages has wide-ranging applicability across a diverse group of disciplines, as the issue of incomplete data is one which plagues researchers of all types.

The main aim of this paper is to develop the methodology for implementation of three-stage multiple imputation. In theory, the imputation stage of implementation is a simple extension of two-stage multiple imputation. However, the combining rules required for drawing appropriate inferences are different and one goal of this paper is to derive the necessary combining rules. Section 2 of this article provides an overview of standard multiple imputation and two-stage multiple imputation practices. Section 3 describes the methodology for implementation of multiple imputation in three stages and provides a proof of the limiting distribution for the proposed estimator. Section 4 gives an extension of the methodology to k-stage multiple imputation. Section 5 defines some common missing data terminology related to mechanisms of missingness and ignorability and presents an extension to the general ignorability assumption for three types of missing values. Section 6 demonstrates the use of multiple imputation in three stages with simulations, showing that the proposed estimator is unbiased and efficient, and Section 7 discusses further extensions of the methodology.

2. Multiple imputation review

2.1. Standard multiple imputation

The idea behind multiple imputation is to fill in plausible values for the missing data several times to account for model uncertainty (Rubin, 1987; Harel and Zhou, 2007). After creating $m > 1$ complete data sets by drawing from the posterior predictive distribution of the missing values, each data set is analyzed using complete data analysis methods. Let Q denote the parameter of interest. An example of such a Q might be a mean or a regression coefficient. From the complete data analyses, complete data estimates ($\hat{Q}$) and their associated variances (U) are obtained.

Let $Y = (Y_{obs}, Y_{mis})$ be the complete data where Y_{obs} is the observed part of the data and Y_{mis} is the missing part of the data. The original derivations for the combining rules were based on large sample inference (Rubin, 1987). Reiter and Raghunathan (2007) review the implications of basing the derivation on large sample inference. The assumption involved was that, in the presence of the complete data, intervals and tests would be based on a normal approximation. That is,

$$(\hat{Q} - Q)/\sqrt{U} \sim N(0, 1).$$

The overall estimate of Q is

$$\bar{Q} = m^{-1} \sum \hat{Q}^{(j)}, \quad j = 1, 2, \ldots, m,$$

where $\hat{Q}^{(j)}$ is the estimate from the jth repeated imputation. To get the variance for $\bar{Q}$, the between-imputation variance and the within-imputation variance must be appropriately combined. The between-imputation variance is denoted by B and is

$$B = (m-1)^{-1} \sum (\hat{Q}^{(j)} - \bar{Q})^2, \quad j = 1, 2, \ldots, m,$$

while the within-imputation variance is denoted by $\bar{U}$ and is

$$\bar{U} = m^{-1} \sum U^{(j)}, \quad j = 1, 2, \ldots, m,$$

where $U^{(j)}$ is the estimated variance of $\hat{Q}^{(j)}$.

The total variance, denoted by T is then equal to

$$T = \bar{U} + (1 + m^{-1})B. \tag{1}$$