Computational Statistics and Data Analysis 82 (2015) 173–185

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern-mixture hot deck

Danielle Sullivan, Rebecca Andridge *

Division of Biostatistics, The Ohio State University College of Public Health, 1841 Neil Ave., Columbus, OH 43210, USA

ARTICLE INFO

Article history:
Received 30 October 2013
Received in revised form 2 September 2014
Accepted 9 September 2014
Available online 19 September 2014

Keywords:
Hot deck
Nonignorable missingness
Donor selection
Sensitivity analysis

ABSTRACT

Hot deck imputation is a common method for handling item nonresponse in surveys, but most implementations assume data are missing at random (MAR). A new hot deck method for imputation of a continuous partially missing outcome variable that harnesses the power of available covariates but does not assume data are MAR is proposed. A parametric model is used to create predicted means for both donors and donees under varying assumptions on the missing data mechanism, ranging from MAR to missing not at random (MNAR). For a given assumption on the missingness mechanism, the predicted means are used to define distances between donors and donees and probabilities of selection proportional to those distances. Multiple imputation using the hot deck is performed to create a set of completed data sets, using an approximate Bayesian bootstrap to ensure "proper" imputations. This new hot deck method creates an intuitive sensitivity analysis where imputations may be performed under MAR and under varying MNAR mechanisms, and the resulting impact on inference can be evaluated. In addition, a donor quality metric is proposed to help identify situations where close matches of donor to donee are not available, which can occur under strong MNAR assumptions. Bias and coverage of estimates from the proposed method are investigated through simulation and the method is applied to estimation of income in the Ohio Medicaid Assessment Survey. Results show that the method performs best when covariates are at least moderately predictive of the partially missing outcome, and without such covariates it effectively reduces to a simple random hot deck for all missingness assumptions.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Incomplete data arise frequently in observational studies, surveys, and even controlled experiments. Inference methods and quality of results are impacted by the amount of missing data and also the reason for the missingness. Consider the simple situation of a single variable, Y, which is subject to missingness, and a single covariate Z which is fully observed. Let M denote the missingness indicator, which takes the value 1 if Y is missing, and 0 if Y is observed. Missing data mechanisms are described by the relationship between M, Y, and Z, and can be classified as one of three types: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) (Rubin, 1987). Under MCAR, $P(M = 1|Y, Z) = P(M = 1)$ and under MAR, $P(M = 1|Y, Z) = P(M = 1|Z)$, neither of which depend on the unobserved values of Y. However, when the probability Y is missing depends upon the missing value itself, i.e., depends on Y, the data

* Corresponding author. Tel.: +1 614 247 7912; fax: +1 614 292 3572.
E-mail addresses: randridge@cph.osu.edu, andridge.1@osu.edu (R. Andridge).

http://dx.doi.org/10.1016/j.csda.2014.09.008
0167-9473/© 2014 Elsevier B.V. All rights reserved.

are MNAR. Tests can be performed to determine if the mechanism is MCAR (Little, 1988a), but if the MCAR assumption is questionable, distinguishing between MAR and MNAR is not possible, and even if the MCAR assumption is not rejected this does not rule out the possibility that data are MNAR.

The missing data mechanism can either be ignorable (meaning one does not need to model it and it can be "ignored"), or nonignorable (meaning that the mechanism should be modeled). An ignorable mechanism results from the data missingness being either MCAR or MAR with distinctness between parameters involved in the data model and the response model. Nonignorability occurs when the data mechanism is either (1) MNAR or (2) MAR and the parameters are not distinct (Rubin, 1987). The assumption of ignorability can be evaluated by subject matter experts, data from outside sources, or more formally through a sensitivity analysis.

There have been many methods developed to handle missing data, all of which range in degree of implementation difficulty, ability to properly reduce bias, and ability to efficiently estimate standard errors. In this paper we focus on one particular method for handling missing data: multiple imputation. Specifically, we develop a nonignorable hot deck imputation procedure that includes a single sensitivity parameter that can be varied in order to assess the impact of possible deviations away from ignorable missingness mechanisms. Throughout we assume that Y is continuous, and, as hot deck procedures are most commonly used to impute one variable at a time, we assume that Y is a single variable subject to missingness with one or more fully observed covariates Z (of any type) available.

The rest of the article is organized as follows. In Section 2 we review two previous approaches to imputation under MNAR that form the basis of our proposed method. In Section 3 we describe the proposed method which we call the proxy pattern-mixture (PPM) hot deck, and in Section 4 we propose a donor quality metric for the hot deck procedure. A simulation study is described in Section 5 to evaluate the performance of the PPM hot deck method compared to alternative approaches and to illustrate the use of the donor quality metric. An application of the PPM hot deck is presented in Section 6 using data from the Ohio Medicaid Assessment Survey (OMAS). Finally, some concluding remarks are presented in Section 7.

2. Background and motivation

Our proposed method combines ideas from two different approaches to multiple imputation under MNAR assumptions that have previously been explored: hot deck imputation using distance-based donor selection and a parametric nonignorable imputation procedure. In this section we first briefly review the use of the hot deck for multiple imputation, and then review the two previously developed methodologies that form the basis for our new imputation method: the nonignorable hot deck of Siddique and Belin (2008b) and the proxy pattern-mixture model of Andridge and Little (2011).

2.1. Hot deck multiple imputation

There are a multitude of versions of hot deck imputation; a review can be found in Andridge and Little (2010). The defining component of hot deck imputation is that for each nonrespondent, a respondent's observation is imputed for the missing value. A simple random hot deck is when all respondents have equal probability of being selected as a "donor" for a missing value. Other versions attempt to find a respondent who is most similar to a nonrespondent, usually based on the values of covariates that are observed for all subjects. A set of possible donors, called a donor pool, is formed for each nonrespondent, and a "close" match is selected out of each donor pool. There are many ways to create the donor pool, numerous metrics used to define a "close" match, and several ways to randomly (or not randomly) select a specific donor. Once a donor has been selected, the donor's observed value is imputed for the nonrespondent's missing value. If a donor is selected randomly from the donor pool, this is referred to as a random hot deck, and this is what we consider in this paper. In the extreme, the "closest" donor is always selected for imputation (i.e., there is not random selection from a donor pool), resulting in the commonly used nearest neighbor imputation method (Chen and Shao, 2000).

The hot deck is often used to create single imputations, which require special methods to properly estimate variances (Burns, 1990; Rao and Shao, 1992; Rao, 1996; Shao and Sitter, 1996; Chen and Shao, 1999). An alternative approach is multiple imputation (MI), proposed by Rubin (1987) as a method to account for the uncertainty associated with missing data, which is the type of imputation we consider in this paper. Instead of imputing a single value for each nonrespondent, multiple imputation results in a set of completed data sets, each containing (possibly different) imputed values for each nonrespondent. If the standard random hot deck procedure is applied multiple times to create a set of imputed data sets, the imputations are not "proper" in that they do not fully propagate variability across imputations. In fact, in Rubin's seminal book (Rubin, 1987) he uses a hot deck MI procedure to illustrate that the between imputation variance will be underestimated if no adjustment is made.

A relatively simple modification to the standard hot deck imputation procedure, called the approximate Bayesian bootstrap (ABB), was introduced by Rubin and Schenker (1991); this modification makes hot deck MI a "proper" MI method. For simplicity, assume there is a single variable, Y, and it is subject to missingness. Let Y_{obs} be a vector of length n_{obs} containing the values of Y that are observed, and Y_{mis} be the vector of length n_{mis} containing the values of Y that are missing. For the purposes of illustration assume there is a single pool of donors, in other words, all n_{obs} subjects with observed Y are eligible to donate to each of the n_{mis} subjects with missing Y. In a standard simple random hot deck, n_{mis} values would be

Download English Version:

https://daneshyari.com/en/article/414965

Download Persian Version:

https://daneshyari.com/article/414965

Daneshyari.com