

# A simple method to ensure plausible multiple imputation for continuous multivariate data

Shakir Hussain, Mohammed Mohammed, Sayeed Haque, Roger Holder, John Macleod

## ▶ To cite this version:

Shakir Hussain, Mohammed Mohammed, Sayeed Haque, Roger Holder, John Macleod. A simple method to ensure plausible multiple imputation for continuous multivariate data. Communications in Statistics - Simulation and Computation, 2010, 39 (09), pp.1779-1784. 10.1080/03610918.2010.518267. hal-00637024

# HAL Id: hal-00637024 https://hal.science/hal-00637024

Submitted on 29 Oct 2011

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



# A simple method to ensure plausible multiple imputation for continuous multivariate data

Journal:	Communications in Statistics - Simulation and Computation
Manuscript ID:	LSSP-2009-0272.R3
Manuscript Type:	Original Paper
Date Submitted by the Author:	06-Aug-2010
Complete List of Authors:	Hussain, Shakir; School of Medicine, University of Birmingham, Division of Primary Care and General Practice Mohammed, Mohammed; University of Birmingham, Department of Public Health Haque, Sayeed; University of Birmingham, Primary Care Holder, Roger; University of Birmingham, Primary Care Macleod, John; University of Bristol
Keywords:	Multiple imputation, implausible imputed values, plausible imputed values
Abstract:	Multiple Imputation (MI) is an established approach for handling missing values. We show that MI for continuous data under the multivariate normal assumption is susceptible to generating implausible values. Our proposed remedy, is to 1) transform the observed data into quantiles of the standard normal distribution, 2) obtain a functional relationship between the observed data and it's corresponding standard normal quantiles, 3) undertake MI using the quantiles produced in step 1 and finally 4) use the functional relationship to transform the imputations into their original domain. In conclusion, our approach safeguards MI from imputing implausible values.



## A simple method to ensure plausible multiple imputation for continuous multivariate data

SHAKIR HUSSAIN<sup>1</sup>, MOHAMMED A MOHAMMED<sup>1</sup>, M. SAYEED HAQUE<sup>2</sup>, ROGER HOLDER<sup>2</sup>, JOHN MACLEOD<sup>3</sup> AND RICHARD HOBBS<sup>2</sup>.

<sup>1</sup>Public Health, Epidemiology and Biostatistics, University of Birmingham, England.
<sup>2</sup>Primary Care Clinical Sciences, University of Birmingham, England
<sup>3</sup>Department of Social Medicine, University of Bristol, Bristol, England.

Multiple Imputation (MI) is an established approach for handling missing values. We show that MI for continuous data under the multivariate normal assumption is susceptible to generating implausible values. Our proposed remedy, is to 1) transform the observed data into quantiles of the standard normal distribution, 2) obtain a functional relationship between the observed data and it's corresponding standard normal quantiles, 3) undertake MI using the quantiles produced in step 1 and finally 4) use the functional relationship to transform the imputations into their original domain. In conclusion, our approach safeguards MI from imputing implausible values.

Keywords: Multiple imputation, implausible imputed values, plausible imputed values

Correspondence to: Dr Shakir Hussain, Senior Statistician. Unit of Public Health, Epidemiology and Biostatistics, University of Birmingham, Edgbaston, Birmingham B15 2TT. E-mail: S.Hussain@Bham.ac.uk.

#### **1. Introduction**

Multiple Imputation (MI) is an established comprehensive approach to dealing with the challenges of missing data. Developed by Rubin (1987) and described further by Schafer (1997) and Little and Rubin (2002), MI methods work by imputing the missing values multiple times and then consolidating across imputed data sets to account for variation within and between imputations reflecting the fact that imputed values are not the known true values. Inevitably MI has to be based on a set of assumptions relating to the distributional form of the variables and the original MI approach generally assumed a joint multivariate normal model for the continuous variables.

Although some researchers have argued that by definition the quality of the imputations cannot be assessed because the missing values are unobserved, there is nevertheless, growing emphasis on the need to assess the quality of the imputations (Abayomi et al. 2008). Amongst the various numerical and graphical approaches suggested to investigate the quality of imputations it seems sensible to ensure that the imputed values are not implausible for the specific application area. For example, the height of a person can only be positive and has a practical upper bound, so imputation of missing heights must also be positive and below the upper bound. The generation of implausible values would suggest problems with the MI which should be remedied before subsequent statistical analyses (Abayomi et al. 2008). Of course if for a particular application the assumption of multivariate normality is not valid then subsequent imputations will be suspect.

In this paper, we use an illustrative data set (household survey data used by Schafer 1997), to demonstrate that MI can sometimes generate implausible values. To safeguard against these implausible imputations we propose a method which essentially makes use of an empirical transformation to normality based on the observed variable in question which naturally constrains any imputed values from MI to fall within the observed range of the variable as well as ensuring the multivariate normality assumption for MI is met.

In Section 2 of the paper, we first illustrate the problem of implausible values arising from the standard application of MI to the household survey data with missing values and show how our proposed strategy mitigates against implausible values. In Section

3 of the paper we offer a practical demonstration of the problem of implausible imputation in practice and apply our approach to data from a healthcare study. In Section 4 we summarise the key issues and conclude the paper.

#### 2. Illustration of implausible imputations

Consider the household survey data included as part of the "norm" package (Novo 2003) in R (R Development Core Team 2010). This small data set (n=25) has five variables (*ageh*, *agew*, *inc*, *edu* and *kid*), where only one variable (*ageh*) is complete. A practical constraint is that no variable can possibly have a negative value. A Kolmogrov-Smirnov test suggested that of the variables with missing data, *edu* and *kid* were not consistent with the normal distribution (test statistic both equal 0.21, p=0.016 and p=0.007 respectively)

We used the missing data library in S-plus version 8.1 (TIBCO 2008) to multiply impute the missing values. We imputed twenty-five data sets on ten occasions and found that on three occasions, the imputed data sets contained one or more negative (implausible) values (see Figure 1) involving the *inc* and/or *edu* variables. Of course, since the negative values were not frequent, the reader may wonder what all the fuss is about, but we show later a data set where all iterations contained implausible values.



Figure 1 Histograms showing the presence of negative values in three of the four imputed variables. Y-axes are counts. Note that each histogram is not necessarily from the same set of imputations.

## 3. Our proposed solution

We now describe our approach to safeguarding MI against negative values without the need to ensure that variables meet the normality assumption. Our steps are as follows:-.

**Step 1:** Transform each raw variable into quantiles of the standard normal distribution. Let  $Y_i$  represent a raw variable in ascending order with some missing values that we are trying to impute. Let  $Z_i$  be the equivalent normal quantile for  $Y_i$ , where

$$Z_i = \Phi^{-1}((i-0.5)/n)$$
 where  $i = 1,...,n$ , and *n* is the sample size of  $Y_i$ 

Note, when  $Y_i$  is missing,  $Z_i$  will also be missing.

Step 2: Derive a functional relationship, such as a second-order polynomial, between  $Y_i$  and  $Z_i$  for non-missing values of  $Y_i$  only. For example

$$Y_{i} = \beta_{0} + \beta_{1} Z_{i} + \beta_{2} Z_{i}^{2} + e_{i}, \qquad (1)$$

where  $\beta_0$ ,  $\beta_1$  and  $\beta_2$  are the coefficients and  $e_i$  is an error term defined to be normally distributed with mean zero and variance  $\sigma^2$ .

**Step 3:** Use MI to derive imputations for the missing in Equation (1) values in  $Z_i$ .

**Step 4:** For each imputed value of  $Z_i$  use Equation (1) to determine its corresponding  $Y_i$ . Substituting a simulated random N(0,  $\sigma^2$ ) value for  $e_i$  rather than its mean value is enables the possible imprecision of the chosen functional relationship to be incorporated.

In the unlikely event that an imputed value of Z, say Z\*, is outside the "empirical" range of  $Z_{i}$  (a plotting position estimate of the first quantile would suggest that the probability of this occurring is approximately 1/n) then we caution against using Equation (1) because we are now extrapolating outside the bounds of the observed data. So, for the special case where Z\* is less than the minimum of  $Z_i$ , all we can say is that the corresponding imputed value of Y, say Y\*, should also be less than the minimum of  $Y_i$ . Similarly, where Z\* is greater than the maximum of  $Z_i$ , Y\* should be greater than the maximum of  $Y_i$ . However if a lower or upper limit to Y was known (zero, for instance) then that might, with caution, be incorporated into the chosen functional relationship to extend its range of validity.

We applied the above procedure to the household data and found no negative values because Equation (1) constrains the imputations to the observed range of the raw variable. We could have incorporated the knowledge that Y must be positive by using a log-polynomial functional relationship.

#### 4. Application to healthcare data

Our motivation stems from data obtained from a follow-up study of the young adult offspring of mothers who participated in a trial of nutritional supplementation during pregnancy. The study aimed to investigate the influence of maternal nutritional status (and other factors) on offspring risk of cardiovascular diseases (Tang et al. 2004). Sixty-five offspring were invited for clinical assessment where measures undertaken included age, gender, body mass index and blood Insulin level - fasting (*If*), thirty minutes (*I30*) and 120 minutes (*I120*) after a standard glucose challenge. Because of incomplete follow-up, 9% of *If* and *I30* and 14% of *I120* were missing. Figure 2 (top row) shows histograms of the three insulin variables. A complete cases only analysis was not deemed to be appropriate because this could lead to biased estimates and loss of precision from a reduced sample size. The pattern of missing data in these variables is an example of monotone missing data and the mechanism was not considered to be missing completely at random (Little's *d*-squared test statistic =3, p=0.08).

Twenty-five multiply imputed datasets were generated, but each set was found to contain one or more negative values (see Figure 2 middle row) - in reality, such values cannot occur. However using our proposed solution we found no negative values (see Figure 2, bottom row) and a Kolmogorov-Smirnov test showed no significant difference between the observed and imputed values using our approach (all test statistics <0.04 with p = 1).



Figure 2: Histograms for the three insulin variables. Top row is observed data. Middle row is the first set of imputed data showing some negative values. Bottom row is the first set of imputed data using our proposed solution showing no negative values.

#### 5. Comment

In this paper we introduce a simple empirical approach to safeguarding MI from imputing implausible values, which we define as being outside the observed range of the variable. We undertook all our analyses using the missing data library in S-plus version 8.1. (TIBCO 2008) and we were also able to replicate our results using the "norm" library (Novo 2003) in R. Our approach is intended to provide some useful guidance but is not prescriptive. There is considerable room for modification. For instance, practitioners may prefer to use *spline* or *lowess* functions or elementary interpolation to describe the functional relationship (see step 2 of our approach) between the observed data and the quantiles from the standard normal.

There are several advantages to our proposed solution. It dovetails into the MI methodology and so can be regarded as pre and post processing around the MI approach thus allowing it to be used as an adjunct to existing MI algorithms. The use of normal quantiles ensures that the multivariate normal assumption is satisfied and there is no reason to suspect that this could adversely interfere with the core MI algorithms. The use of quantiles is also well suited to ordinal variables which are frequently met in practice. Whilst it may be argued that transforming the raw variable to achieve normality can mitigate against the production of implausible imputed values, our illustrative example shows that, even when variables are apparently consistent with the normal distribution, MI is still susceptible to the production of implausible values. Actually, our approach obviates the need to transform the raw variable to normality, which is useful because some variables will, even after transformation, not meet the normality assumption adequately. Our approach seeks to ensure that an imputed value does not fall outside of the range of the observed data. This concurs with the advice to exercise caution when extrapolating outside the range of the observed data, but seems unsatisfactory for situations where the missing value is somehow known to lie outside the range of the observed values. We suggest that where the feasible range of the variable is known this may, with caution, be incorporated into the functional relationship chosen to extend its range of validity.

Interestingly, we imputed our healthcare data in two more recent R libraries – "mi" (Gelman et al. 2009) and "mice" (van Buuren and Groothuis-Oudshoorn 2009). Fortunately, neither produced implausible values. Since the S-plus missing data

library (and "norm" in R) relies on a joint multivariate normal assumption and "mi" and "mice" are based on more recent formulations involving multiple imputation by chained equations (Raghunathan et al. 2001), it would seem that the former specification may be more susceptible to implausible values although the latter specification is not without its own challenges (Stuart et al. 2009). Nevertheless the general question of implausibility in either MI paradigm needs greater emphasis and further research (Abayomi et al. 2008, Stuart et al. 2009). Meanwhile we suggest that, at least for MI based on the multivariate normal assumption, our empirical approach offers a simple way to safeguard against implausible imputed values.

Acknowledgments: We are grateful to the anonymous reviewers for their helpful criticism and comments on earlier drafts of this paper.

## **References**

Abayomi, K., Gelman, A., Levy, M. (2008). Diagnostics for Multivariate Imputations. *Journal of the Royal Statistical Society Series C Applied Statistic* 57:273-291.

Gelman, A., Hill, J., Su, Y., Yajima, M., Pittau, M. G. (2009). The mi package. Available at: http://cran.r-project.org/web/packages/mi/mi.pdf

Little, R. J. A., Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Second edition. New York Wiley.

R Development Core Team. (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Austria. Available at: http://www.R-project.org.

Raghunathan, T.E., Lepkowski J.M., van Hoewyk J., Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 27:85–95.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York Wiley.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall/CRC.

Stuart, E., Azur, M., Frangakis, C., Leaf, P. (2009). Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. *American Journal of Epidemiology* 169:1133-1139.

Tang, L.P., MacLeod, J.A., Hobbs, F.D.R.H., Wharton, B.A., Davey Smith, G., Stewart, P.M. (2004). Fetal origins of adult disease: tracing and recruitment of offspring whose mothers participated in a trial of nutritional supplementation during pregnancy – the Sorrento experience. *Nutrition Bulletin* 29:310-316.

TIBCO Software Inc (2008). Analyzing data with missing values in TIBCO Spotfire

<text>