International Journal of Education & Literacy Studies
ISSN 2202-9478
Vol. 2 No. 3; July 2014

Standard and Robust Methods in Regression Imputation

Behjat Moraveji
Department of Basic Science
Fars Science and Research Branch
Islamic Azad University, Fars, Iran

Koorosh Jafarian
Department of Foreign Languages,
Bandar Abbas Branch
Islamic Azad University, Bandar Abbas, Iran
E-mail:koorosh_jafarian2001@yahoo.com

Received: 10-05- 2013 Accepted: 20-06-2013 Published: 30-07-2014
doi:10.7575/aiac.ijels.v.2n.3p.32 URL: http://dx.doi.org/10.7575/aiac.ijels.v.2n.3p.32

Abstract

The aim of this paper is to provide an introduction of new imputation algorithms for estimating missing values from official statistics in larger data sets of data pre-processing , or outliers. The goal is to propose a new algorithm called IRMI (iterative robust model-based imputation).This algorithm is able to deal with all challenges like representative and non-representative outliers and a mixture of different distributions of variables. This algorithm is compared to the algorithm IVEWARE to illuminate the advantages and disadvantages of different techniques for imputation in artificial data and real data sets from official statistics , with respect to robustness are proposed, especially in presence of outliers the model-based of new algorithm is preferable.

Keywords: IRMI, IVEWARE, imputation, robustness

1. Introduction

From the last few decades to now ,empirical researchers in their data sets often are confronted with missing values , and in the last years many developments have been made to increase computing power. The imputation of the missing values in real statistic is necessary. Although in statistical literature on the estimation of missing values we have two types of non responses such as unit non response and item non response, data sets often contain missing values, and they must be replaced by meaningful values .The meaningful values for estimation in missing data is known with the name of imputation (Little & Rubin, 1987). For using a proper imputation method one must be known the missing data mechanism(s). The quality and quantity of the imputed values depends on the imputation itself and on the imputation method .

The objective of imputation can be summarized in two cases (Rubin, 1987):

(1) Instructions and data outputs to allow users to use standard analytical tools for data containing

missing values;
(2) providing statistically valid inferences for data containing missed by applying imputation methods.

2. Imputation methods

1. The detective of the missing values exactly is difficult in practice, because this case requires the knowledge of the missing value (Little & Rubin, 1987). Many different methods for the estimation of meaningful values in imputation have been developed in the last decades. They may be divided into univariate (single) methods such as mean imputation, and multivariate (multiple) imputation(MI)-introduced by Rubin(1987). In the latter case there are basically several approaches: model- based imputation methods such as regression imputation or k-nearest neighbor imputation, covariance methods such as the approaches by Verboven , Branden, and Goos (2007) or Serneels and Verdonck (2008), and model-based imputation methods such as (EM-based) regression imputation.

2. Item non response for imputation is searched by using iterative model-based imputation methods. Several strategies are possible to choose the non response (random or stochastic imputation, deterministic imputation, etc.). To decrease the computation time and variance for imputation large data sets, the implementation is done by different imputation like: simple deterministic ,model-based deterministic ,simple random and model-based random imputation. In addition, data sets consist of variables that have different distribution, i.e. a variable could be categorical (nominal and ordered variables) to be continuous distribution and the other variables (binary and etc.) to be semi-continuous distributed(see. e.g. Schafer & Olson,1999).

3. If an imputation method of missing values is proper (see, e.g., Rubin 1987), we will be able to apply multiple imputation(MI), generating more than one candidate for each missing cell. However, the sampling variability can be affected by adding the level of noise to the imputed meaningful values by applying bootstrap methods (Little & Rubin 1987, Alfons, Templ,& Filzmoser 2009).For example, In this section a simulation study for the purpose of illustration is provided. Raessler and Munnich (2004)give a description on how to use simulation. Assume that age (AGE) is normally distributed with mean 60 [years] and standard error of 10 [years], income (INC) as normally distributed with mean 1700 [EURO] and standard error of 400 [EURO]. Moreover, let the correlation between age and income be about 0.9. So we let

$$(\text{AGE, INC}) \sim N\left(\begin{pmatrix} 60 \\ 1700 \end{pmatrix}, \begin{pmatrix} 10^2 & 09.300 \\ 0.9.3000 & 400^2 \end{pmatrix}\right)$$

A sample of n = 3000 is drawn. After being generated, the AGE variable into 6 categories is ranged, 1 <= 20 years, 2 = 20 - 30 years, ..., 6 > 60 years. First, the complete cases are analyzed, the mean income estimate, its standard error (s.e.), and the 95% confidence interval are calculated. Then different missingness mechanisms (MCAR, MAR, MNAR,MBND) are applied on income. Under MAR, income is missing with higher probability when age is higher, under MNAR, the probability missing in income is higher for higher itself.

Table 1. verifies how precision is reduced when only the cases are used under MCAR,MAR,MNAR,MBND and mean imputation in IVEWARE,IMI,IRMI-MM.

Table 1. Result of the simulation study

Missing	mPop	CA	Mean	IVEWARE	IMI	IRMI-MM
none	0.97					
MCAR		0.965	0.817	0.913	0.918	0.902
MNR		0.711	0.539	0.906	0.849	0.885
MNAR		0.822	0.845	0.920	0.961	0.900
MBND		0.796	0.575	0.917	0.918	0.858

4. Model based methods without robustness propose that the data originate from a multivariate normal distribution (for instance, the MCMC methods of the imputation software MICE in van Buuren &Oudehoorn (2005), Amelia (Honaker, King &Blackwell, 2009), mi (Yu Sung, Gelman, Hill &Yajima,2009) or IVEWARE (Raghunathan, Lepkowski &Hoewyk 2001). This proposal becomes inappropriate in the presence of outliers in the data, or in case of skewed or multimodal distributions and such challenges in the data of almost all real world data sets. We propose imputation methods based on robust estimates should be applied. These methods proximally give the same result when the data originate from a multivariate normal distribution, and give reliable estimates.

5. The last procedure behind most model-based imputation methods is the EM-algorithm (Dempster,Laird, & Rubin,1977) ,which can be applied for the application of iterative estimation, adaption and re-estimation. For the estimation ,we usually apply regression methods in an iterative manner, which is known under the names regression switching, chain equations, sequential regressions, or variable-by-variable Gibbs sampling (see, e.g.,van Buuren & Oudshoorn 2005, Muennich & Rassler , 2004).

Although, Schafer (2009, 1997), Schafer and Olson (1999) described the problems for semi-continuous variables as suitable for multiple imputation in general, it also has the same limitations. This problem also exists in the algorithm IVEWARE. These algorithms in mi and IVEWARE are based on iterative regression imputation.multiple imputation (mi) initiates the algorithm by a rough initialization (randomly chosen values). The same idea is used by MICE (Multiple imputation by chain equation) by package of van Buuren and Oudshoorn (2005), and the Amelia package of Honaker et al. (2009).

All above algorithms and procedures cannot cope with representative of outliers. The goal is to develop a procedure that is better than the above algorithms, but has the additional feature called robustness for data outliers. Since IVEWARE has all the mentioned problems except robustness. it is natural to use for our task.

3. Missing Values Mechanisms

There are four important terminologies introduced by Little and Rubin (1987, 2002), and Schafer(1996) . Missing values mechanisms can be classified according to the responses.The data set is composed of two components, complete data and missing data (X_miss, X_obs) , X_obs :variable that all values are completely observed and X_ miss : variable that has some missing values.

1.MCAR(the distribution of missingness neither depends on the observed part Xobs nor on the missing part Xmiss,the missing value are said to Missing Completely At Random (MCAR). Thus the probability of missingness is given by P(Xmiss|X) = P(Xmiss) with the complete data X = (Xobs,Xmiss).

2.MAR(the distribution of missingness depends on the observed part Xobs, the missing values are said to be Missing At Random (MAR), and the probability of missingness is P(Xmiss|X) = P(Xmiss|Xobs).

3.MNAR(the distribution of missingness depends on Xmiss, the missing data are said to be Missing Not At Random (MNAR). Thus the probability is given by P(Xmiss|X) = P(Xmiss|(Xobs,Xmiss)).

Hence, the missings can not be fully explained by the observed part of the data.

4. MBND(the distribution of missing value depends on natural design in order not to measure it , the missing values are said to be Missing By Natural Design(MBND).

4. The Algorithm IVEWARE

This algorithm *IVEware* is a software application that is built on the SAS macro Language and a set of independent C and FORTRAN routines and can perform:

1. Single or multiple imputations of missing values using the Sequential Regression Imputation Method (*Survey Methodology*, June 2001).

2. A variety of descriptive and model based analyses accounting for such complex design features as clustering, stratification and weighting .

3. Imputation analyses for both descriptive and model-based survey statistics.

IVEware includes four modules: IMPUTE, DESCRIBE, REGRESS and SASMOD□

(1)IMPUTE: a multivariate sequential regression approach to imputing item missing values.

(2)DESCRIBE**:** the estimation of the population means, proportions, subgroup differences, contrasts and linear combinations of means and proportions.

(3)REGRESS: the fitness of the linear, logistic, polytomous, Poisson, to bit and proportional hazard regression models for data resulting from a complex sample design.

(4)SASMOD: users allow to take into account complex sample design features when analyzing data with several SAS procedures. Currently the following SAS PROCS can be called: CALIS, CATMOD, GENMOD, LIFEREG, MIXED, NLIN, PHREG, and PROBIT.

The algorithm IVEWARE fits a sequence of regression models for the estimation of the missing values and drawing values from the corresponding predictive distributions (Raghunathan et al., 2001). In few words, the algorithm includes the following steps:

(1) Sort the variables according to the amount of missing values.

(2) Initialization loop: Initialize the missing values by using one variable with missing values as response and the variables which include either no missing values or variables which are already initialized as predictors. Apply this procedure for all variables which includes missing values.

(3) Iteration: Use one variable as response and the others as predictors and update the missing values in the response by drawing from the predictive distribution. The link function of the generalized linear regression method has to be selected based on the distribution of the response. Do that for all variables. Repeat this procedure until convergency**.**

5. The Algorithm IRMI

Templ and Filzmoser (2008) and Templ, Kowarik, and Filzmoser (2008) offer the description of new algorithm called IRMI which stands for iterative robust model-based imputation. The corresponding function is called function irmi . It mimics the functionality of IVEWARE (Raghunathan, Lepkowski, & Hoewyk 2001), but there are some advantages such as robustness properties for the estimation procedures. In each step of the iteration, one variable is used as a response variable and the remaining variables serve as the regressors and the multivariate information will be used for imputation in the response variable. The procedure algorithm can be summarized as follows:

(1) Initializes the missing values.

(2) Selects one variable as response and the remaining variables as predictors, and updates the former missing values in the response.

(3) Goes to the next variable and repeats the procedure.

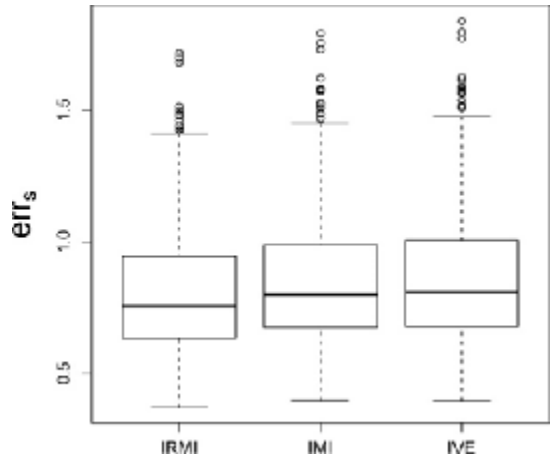
(4) Repeats the whole procedure which is started from 2 until convergence.

(5)Adds noise to final estimates in the a proper way to allow for multiple imputation.

It should be noted that robust regression protects against poorly initialized missing values. Function irmi also shows better behavior for real-world data in practice. All these data sets contain missing values such as the European Union Statistics of Income and Living Conditions (EU-SILC) survey 2004 from Statistics Austria for measuring poverty and social classes in Europe, the Austrian structural business statistics data (SBS) from 2006, a census data set from 1994 provided by the University of California (for details, see http:// www.ics.uci.edu /~mlearn /MLRepository.html),and daily air quality measurements in New York, May to September 1973 (see also Chambers ,Cleveland , Kleiner, & Tukey 2008).

6. Application to Real Data

For comparing the imputation algorithms in the literature on data set from official statistic, we use a data set which contains missing values. In spite of this simplification, the result should reflect the performance of the algorithms for the data. Data cells are randomly selected and set to missing. The imputed values are compared with their true original values after imputation(e.g. daily air quality measurements in New York,from May to September 1973 (see also Chambers et al,.2008). It consist of 200 observations on 5 variables (solar, wind, temp, month, day),and the first two variables contain missing values. Only the complete observations (150) and the first 4 variables are used. Missing values are set in the first two variables according to the proportion in the original data set.



Result of above figure shows that IRMI gives the best result.

7. Conclusions

In general, almost all real-world data sets, especially in official statistics, include outlying observations so that they include different types of distributions. The estimation of missingness in multivariate data can be done better function rather than univariate imputation methods. While, some procedures are depends on the estimation of the multivariate data structures. We summarized an iterative robust model-based imputation procedure for imputation of missing values ,which can deal with the mentioned data problems. In principle, this algorithm is robust against outliers. All simulation ,artificial and real data results show that our robust method shows equal behavior or outperforms the investigated non-robust methods. In addition to, the results from the imputation of the realistic data and popular air quality data set which includes several semi-continuous variables showed that IRMI performs very well in a real-world settings. Therefore, we would suggest to apply IRMI when the variables are include different types of distributions possible such as, binary, categorical, semi-continuous and continuous variables.

References

Alfons, A., Templ, M., & Filzmoser, P. (2009). *On the influence of imputation methods on laeken indicators: Simulations and recommendations*. In: UNECE Work Session on Statistical Data Editing; Neuchatel, Switzerland. p.10URLhttp : // www.unece.org/stats/document/ece/ces/eg.44/2009/ wp.36.e. pdf.

Chambers,J., Cleveland, W.,Kleiner, B., & Tukey,P. (2008).“*Graphical methods for Data Analysis*”. Wadsworth, Belmont, CA.

Dempster, A., Laird, N., and Rubin, D. (1997). *Maximum likelihood for incomplete via the EM algorithm (with discussion)*. Journal of the Royal Statistical Society . 39, 1-38.

Honaker, J., King, G., & Blackwell, M. (2009). *Amelia: Amelia II: A Program for Missing Data*. R package version 1.2-2. URL http://CRAN.R-project.org/package=Amelia.

Little, R., Rubin, D. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

Muennich, R., & Rassler, S. (2004). *Variance estimation under multiple imputation*.In: Proceedings of Q2004 European Conference on Quality in Survey Statistics , Mainz. p. 19.

Raessler, S., & Munnich, R. (2004).*The impact of multiple imputation for DACSEIS”*.Research report ist-2000-26057-dacseis, 5/2004, University of Tubingen.

Raghunathan, T., Lepkowski, J., Hoewyk, J., (2001). *A multivariate technique for multiply imputing missing values using a sequence of regression models*.Survey Methodology 27 (1), 85-95.

R Development Core Team.R. (2009).*A language and Enviroment for Statistical Computing*.R Foundation for Statistical Computing .Vienna,Austria.ISBN 3-900051-07-0.URl:http://www.R-project.org

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

Schafer, J. (1996). *Analysis of Incomplete Multivariate Data*. Chapman & Hall,chapter 9.

Schafer, J. (2009). *Mix:Estimation/multiple Imputation for Mixed Categorical and Continuous Data*". R package version 1.0-7, see also his implementation in SAS and SPLUS. URL http://CRAN.R-project.org/package=mix.

Schafer, J., & Olson, M. (1999).*Modeling and imputation of semicontinuous survey variables".* Fcsm research conference papers, Federal Commitee on Statistical Methodology. URL http://www.fcsm.gov/99papers/shaffcsm.pdf

Serneels,S.,Verdonck,T.,(2008).*Principal component analysis for data containing outliers and missing elements*. Computational Statistics & Data Analysis.52(3),1712_1727.

Templ,M., Alfons, A., & Kowarik, A. (2009).*VIM:Visualization and Imputation of Missing Values*. R package version 1.2.4. URL http://cran.r-project.org.

Templ,M., and Filzmoser, P.,(2008). *Visualization of missing value using the R-package VIM*. Research Templ, M.,and Filzmoser,P.(2008) ."*Visualization of missing value using the R-package VIM*. Research Report cs-2008-1 .Department of Statistics and Probability Theory ,Vienna of Technology . URI http: //www.statistik .tuwein. ac . at. /forschung / CS / CS -2008-1 complete .pdf.

Van Buuren, S ., and Oudshoom , C. (2005).*Flexible Multivariate Imputation by MICE.* Tno/vgz/pg99.054 , Netherlands Organization for Applied Scientific Research (TNO). URI http: //web.inter.nl.net?users/S.van.Buuren/mi/docs/rappory99054.pdf.

Verboven, S., Branden, K., Goos, P.,(2007). *Sequential imputation for missing values.* Computational Biology and Chemistry 31, 320_327.

Yu-Sung,S.,Gelman, A., Hill, J., & Yajima, M., (2009). Multiple imputation with diagnostics (mi) in R :opening windows into the black box. *Journal of Statistical Software.*