A C T A   U N I V E R S I T A T I S   L O D Z I E N S I S
FOLIA OECONOMICA 269, 2012

Małgorzata Misztal[*]

IMPUTATION OF MISSING DATA USING R PACKAGE

Abstract. Missing data are quite common in practical applications of statistical methods. Imputation is general statistical method for the analysis of incomplete data sets.

The goal of the paper is to review selected imputation techniques. Special attention is paid to methods implemented in some packages working in the R environment. An example is presented to show how to handle missing values using a few procedures of single and multiple imputation implemented in R.

Key words: missing values, single imputation, multiple imputation, R – project.

I. INTRODUCTION

Incomplete data are quite common in practical applications of statistical methods. Dealing with data sets with missing values researchers often discard observations with any missing values and perform complete case analysis. It can lead to biased estimates, incorrect standard errors and incorrect inferences or results.

Another way to deal with missing data is to impute all missing values before analysis, using single or multiple imputation methods.

The goal of the paper is to review selected imputation techniques implemented in some packages working in the R environment. An example is presented to show how to handle missing values using different imputation methods implemented in R.

II. IMPUTATION PROCEDURES

Using any method of dealing with missing values it is important to understand why the data are missing. Little and Rubin (2002) described three missing data mechanisms: *missing completely at random* (MCAR), *missing at random* (MAR) and *not missing at random* (NMAR).

According to Molenberghs and Kenward (2007, p. 4), the MCAR mechanism potentially depends on observed covariates, but not on observed or unobserved

[*] Ph.D., Chair of Statistical Methods, University of Łódź.

outcomes. The MAR mechanism depends on the observed outcomes and perhaps also on the covariates but not on unobserved measurements. Finally, the NMAR mechanism depends on unobserved measurements perhaps in addition to dependencies on covariates and on observed outcomes.

For MCAR mechanism the observed values are essentially a random sample of the full data set so the complete case analysis gives the same results as the full data set would have.

Under an assumption of MAR mechanism handling missing data one can use (among others) *imputation – based procedures* or *model - based* ones.

In *imputation – based techniques* the missing values are filled in (using single or multiple imputation methods) and the complete data are analyzed by standard statistical methods. Some details are listed below.

In *model – based procedures* one should define a model for the observed data – inferences are based on the likelihood or posterior distribution under that model with parameters estimated by procedures such as maximum likelihood – see Little and Rubin (2002) for details.

Since imputations are means or draws from a predictive distribution of the missing values, there is a need for a method creating such a predictive distribution for the imputation based on the observed data. Little and Rubin (2002) state that there are two approaches to generating this distribution:

1. *Explicit modeling* – where the predictive distribution is based on a formal statistical model (e. g. multivariate normal);
2. *Implicit modeling* – where the focus is on the algorithm, which implies an underlying model.

The most popular explicit modeling methods are:

(1) *mean /mode imputation* – for any continuous variable missing values are imputed using the mean of the observed values; for categorical variables the mode is used;

(2) *conditional mean imputation* (*regression imputation*) – missing values are replaced by predicted values from a regression model relating predictor with missing values to all other predictors; least squares, logistic and ordinal regressions are used with continuous, binary and ordered categorical predictors, respectively.

(3) *stochastic regression imputation* – missing values are imputed by predicted values from a regression model plus a residual.

The most popular implicit modeling methods are:

(1) *hot deck imputation* – missing values are imputed using sampling with replacement from the observed data;

(2) *substitution* – nonresponding units are replaced with alternative units not selected into the sample;

(3) *cold deck imputation* – missing values are filled in by a constant value from an external source;

(4) *predictive mean matching* – combination of regression imputation and hot deck method – the method starts with regressing the variable to be imputed – Y - on a set of predictors for cases with complete data; on the basis of this regression model predicted values are generated for both the missing and non-missing cases; then for each case with missing data, a set of cases with complete data that have predicted values of Y that are "close" to the predicted values for the case with missing data is found and from this set of cases one is randomly chosen – its Y value is used to impute the missing case (see Allison 2002).

Single imputation does not take into account the uncertainty in the imputations. That's why *multiple imputation* (MI) is recommended as appropriate way of handling missingness in data. There are three steps of multiple imputation process (Yu et al. 2007):

I. generate m>1 imputed data sets by filling in the missing values with plausible values;

II. perform standard analyses on each of the m imputed data sets;

III. combine the results from the m analyses.

According to van Buuren and Groothuis-Oudshoorn (2010) there are two general approaches to multiple imputation: *joint modeling* (JM) proposed by Schafer (1997) and *fully conditional specification* (FCS) developed by van Buuren (2007).

Joint modeling entails specifying a multivariate distribution for the missing data and drawing imputation from their conditional distributions by Markov Chain Monte Carlo (MCMC) techniques (e.g. data augmentation).

FCS is based on the iterative process that involves specifying a conditional distribution for each incomplete variable. It does not explicitly assume a particular multivariate distribution, but assumes that one exists and draws can be generated from it using Gibbs sampling (see Yu et al. 2007). The imputed values can be either the predicted values sampled from the posterior distribution of the incomplete variable or obtained using predictive mean matching as the observed value from the complete case with the closest predicted value to the incomplete case.

MCAR and MAR mechanisms are called *ignorable* ones and there are a lot of techniques for handling ignorable missing data.

NMAR mechanism is called *non-ignorable* and requires a different and more complex approach, i. e. selection models or pattern-mixture models (see details in Allison 2002, Little and Rubin 2002 or Molenberghs and Kenward 2007).

III. IMPUTATION SOFTWARE

Imputation techniques are implemented in some statistical packages. SOLAS (Statistical Solutions Inc, Sargus, MA, USA) is a specific software package designed for handling missing data and performing multiple imputations.

Several standard statistical packages – SAS, SPSS, STATA and R-project have also implemented standard and user – written programs for dealing with missing data. The performances of these packages are compared for example by Yu et al. (2007) or by Horton and Kleiman (2007). In this paper only R-project is taken under consideration.

In R missing values are indicated by NA's. There are (at least) 11 packages, working in the R environment, to handle missing data: Amelia II, Hmisc, mi, mice, yaImpute, mix, cat, norm, pan, monoman, mvnmle. Another two packages – mitools and VIM can be useful to combine the results from multiple imputations and to explore the data and the structure of the missing values. Short description of every package is presented in Table 1.

Some of the packages mentioned above are used in an example.

IV. EXAMPLE

Let's consider the data set of 467 people that were granted a consumer credit. The aim of the study was to classify the borrowers into two risk classes: bad (defaulted loans) and good (paid off loans).

There were 6 independent variables (age, loan amount, borrower's seniority in months, average income of the last three months, monthly installment, loan period in months). Decision rules were established on the basis of logistic regression model.

From the complete data set of 467 objects, 5.72% of values were randomly removed and replaces by NA's.

Data with missing values are stored in the cred.txt file and read into R using the command:

```
> cred=read.table("C:/Documents and Settings/dane/cred.txt",
header=TRUE).
```

Using logistic regression model with the complete original data set produces the results presented in Table 2.

Discarding observations with any missing value there are 294 cases for complete case analysis. The results from complete case analysis using logistic regression are also summarized in Table 2. The Design package was used to estimate the logistic regression model coefficients.

Table 1. Handling missing data with R – basic information

Package	Version/ Date	Title	Authors	Description	Basic command
1	2	3	4	5	6
Amelia II	1.2-18 2010-11-04	Amelia II: A Program for Missing Data	James Honaker, Gary King, Matthew Blackwell - Harvard University	Uses a bootstrap+EM algorithm to impute missing values from a dataset and produces multiple output datasets for analysis	amelia(x, m = 5, p2s = 1, frontend = FALSE, idvars = NULL, ts = NULL, cs = NULL, polytime = NULL, splinetime = NULL, intercs = FALSE, lags = NULL, leads = NULL, startvals = 0, tolerance = 0.0001, logs = NULL, sqrts = NULL, lgstc = NULL, noms = NULL, ords = NULL, incheck = TRUE, collect = FALSE, arglist = NULL, empri = NULL, priors = NULL, autopri = 0.05, emburn = c(0,0), bounds = NULL, max.resample = 100, ...)
Hmisc	3.8-3 2010-09-08	Harrell Miscellaneous	Frank E Harrell Jr - Vanderbilt University School of Medicine	Multiple Imputation using Additive Regression, Bootstrapping, and Predictive Mean Matching	aregImpute(formula, data, subset, n.impute=5, group=NULL, nk=3, tlinear=TRUE, ype=c('pmm','regression'), match=c('weighted','closest'), fweighted=0.2, curtail=TRUE, boot.method=c('simple', 'approximate bayesian'), burnin=3, x=FALSE, pr=TRUE, plotTrans=FALSE, tolerance=NULL, B=75)
				Transformations/Imputations using Canonical Variates	transcan(x, method=c("canonical","pc"), categorical=NULL, asis=NULL, nk, imputed=FALSE, n.impute, boot.method=c('approximate bayesian', 'simple'), trantab=FALSE, transformed=FALSE, impcat=c("score", "multinom", "rpart", "tree"), mincut=40, inverse=c('linearInterp','sample'), tolInverse=.05, pr=TRUE, pl=TRUE, allpl=FALSE, show.na=TRUE, imputed.actual=c('none','datadensity','hist','qq','ecdf'), iter.max=50, eps=.1, curtail=TRUE, imp.con=FALSE, shrink=FALSE, init.cat="mode", nres=if(boot.method=='simple')200 else 400, data, subset, na.action, treeinfo=FALSE, rhsImp=c('mean','random'), details.impcat='', ...)

Table 1 (cont.)

1	2	3	4	5	6
mice	2.4 2010-10-18	Multivariate Imputation by Chained Equations	Stef van Buuren (TNO Quality of Life, Leiden + University of Utrecht) & Karin Groothuis-Oudshoorn (Roessingh RD, Enschede + University Twente)	Multiple Imputation using Fully Conditional Specification	mice(data, m = 5, method = vector("character",length=ncol(data)), predictorMatrix = (1 - diag(1, ncol(data))), visitSequence = (1:ncol(data))[apply(is.na(data),2,any)], post = vector("character", length = ncol(data)), defaultMethod = c("pmm","logreg","polyreg"), maxit = 5, diagnostics = TRUE, printFlag = TRUE, seed = NA, imputationMethod = NULL, defaultImputationMethod = NULL)
mi	0.09-11.03 2010-11-11	Missing Data Imputation and Model Checking	Andrew Gelman, Jennifer Hill, Yu-Sung Su, Masanao Yajima, Maria Grazia Pittau - Columbia University	Multiple Iterative Regression Imputation – the basic command generates a multiply imputed matrix applying the elementary functions iteratively to the variables with missingness in the data randomly imputing each variable and looping through until approximate convergence	mi(object, info, n.imp = 3, n.iter = 30, R.hat = 1.1, max.minutes = 20, rand.imp.method = "bootstrap", run.past.convergence = FALSE, seed = NA, check.coef.convergence = FALSE, add.noise = noise.control())
yaImpute	1.0-12 2010-09-01	yaImpute: An R Package for k-NN Imputation	Nicholas L. Crookston & Andrew O. Finley - Michigan State University	Performs popular nearest neighbor routines for imputation	Find K nearest neighbors: yai(x=NULL, y=NULL, data=NULL, k=1, noTrgs=FALSE, noRefs=FALSE, nVec=NULL, pVal=.05, method="msn", ann=TRUE, mtry=NULL, ntree=500, rfMode="buildClasses") Impute variables from references to targets: impute(object, ancillaryData=NULL, method="closest", method.factor=method, k=NULL,vars=NULL, observed=TRUE,...)

Table 1 (cont.)

1	2	3	4	5	6
mix	1.0-8 2010-01-03	Estimation/multiple Imputation for Mixed Categorical and Continuous Data	Joseph L. Schafer - The Pennsylvania State University	Imputes Missing Data Under General Location Model	imp.mix(s, theta, x)
norm	1.0-9.2 2010-04-29	Analysis of multivariate normal datasets with missing values	Ported to R by Alvaro A. Novo. Original by Joseph L. Schafer	Imputes missing multivariate normal data	imp.norm(s, theta, x)
cat	0.0-6.2 2009-07-28	Analysis of categorical-variable datasets with missing values	Ported to R by Ted Harding and Fernando Tusell. Original by Joseph L. Schafer	Imputes missing categorical data -performs single random imputation of missing values in a categorical dataset under a user-supplied value of the underlying cell probabilities	imp.cat(s, theta)
pan	0.2-6 2009-04-19	Multiple imputation for multivariate panel or clustered data	Joseph L. Schafer - The Pennsylvania State University	Imputation of multivariate panel or cluster data using the Gibbs sampler algorithm	pan(y, subj, pred, xcol, zcol, prior, seed, iter=1, start)
monomvn	1.8-3 2010-04-23	Estimation for multivariate normal and Student-t data with monotone missingness	Robert B. Gramacy – University of Chicago	Maximum likelihood estimation of the mean and covariance matrix of multivariate normal (MVN) distributed data with a monotone missingness pattern	monomvn(y, pre = TRUE, method = c("plsr", "pcr", "lasso", "lar", "forward.stagewise", "stepwise", "ridge", "factor"), p = 0.9, ncomp.max = Inf, batch = TRUE, validation = c("CV", "LOO", "Cp"), obs = FALSE, verb = 0, quiet = TRUE)

Table 1 (cont.)

1	2	3	4	5	6
mvnmle	0.1-8 2009-04-17	ML estimation for multivariate normal data with missing values	Kevin Gross, with help from Douglas Bates, North Carolina State University	Finds the maximum likelihood estimate of the mean vector and variance-covariance matrix for multivariate normal data with missing values	mlest(data, ...)
mitools	2.0.1 2010-05-07	Tools for multiple imputation of missing data	Thomas Lumley – University of Auckland	Tools to perform analyses and combine results from multiple-imputation datasets	MIcombine(results,variances,call=sys.call(), df.complete=Inf,...)
VIM	1.4.2 2010-10-20		Matthias Templ, Andreas Alfons, Alexander Kowarik - Vienna University of Technology	Package introduces new tools for the visualization of missing values in R, which can be used for exploring the data and the structure of the missing values	A lot of commands for visualization and exploring missing data

Source: Self-prepared on the basis of *Manuals* available on http://www.r-project.org/.

Table 2. The results of using logistic regression model
– original data and complete case analysis.

Complete original data set (no missing values, n=467)				Complete case analysis (no missing values, n=264)			
Variables	Coeff.	SE	p-value	Variables	Coeff.	SE	p-value
Intercept	-0.15900	0.5967	0.7899	Intercept	-0.36000	0.8438	0.6696
X1	-0.01883	0.0110	0.0855	X1	-0.00443	0.0137	0.7466
X2	-0.00004	0.0002	0.8060	X2	-0.00009	0.0003	0.7660
X3	-0.00507	0.0017	**0.0036**	X3	-0.00549	0.0029	0.0612
X4	-0.00059	0.0002	**0.0006**	X4	-0.00066	0.0002	**0.0045**
X5	0.00392	0.0026	0.1282	X5	0.00343	0.0043	0.4273
X6	0.05681	0.0220	**0.0097**	X6	0.04718	0.0315	0.1346

Source: Author's calculations.

The results of fitting the logistic regression model to some data sets obtained from using different strategies for dealing with missing data are summarized in Table 3. Seven procedures were employed for handling missing data. A short description and some examples of commands in R are presented below.

The most popular and often used in practice single imputation method is *mean imputation* – for each continuous predictor missing values are imputed using the mean of the observed values. Assuming that mean imputed complete data set is denoted as cred_mean.txt the following list of commands gives the set of estimated regression coefficients and fitted probabilities:

```
> require(Design)
> cred_mean=read.table("C:/Documents and Settings/dane/cred_mean.txt", header=TRUE)
> cred.mean.lr=lrm(Y~X1+X2+X3+X4+X5+X6, data=cred_mean, method="lrm.fit")
> cred.mean.lr
> cred.mean.pred=predict.lrm(cred.mean.lr, type="fitted")
```

The next single imputation method used in the example is the *nearest neighbor search and imputation* procedure, implemented in the yaImpute package. The complete data set can be obtained with the following commands:

```
> require(yaImpute)
> x=as.data.frame(cred[,"Y"]) # the list of variables measured on all observations
> y=cred[, c("X1", "X2", "X3", "X4", "X5", "X6")] # the list of variables with missing values
> cred.yai=yai(x=x, y=y, data=cred, method="euclidean") # the kNN search
> cred.yai.imp=impute(cred.yai) # imputation
```

Since single imputation methods suffer from the problem that tests and confidence intervals are distorted by overstated precision, multiple imputation procedures have been developed to alleviate this problem (Ambler et al. 2007). Three packages working in the R environment are used in our example: Amelia II, Hmisc and mice.

Multiple imputation using the Amelia package can be made using the following list of commands:

```
> require(Amelia)
> bds=matrix(c(2,3,4,5,6,7,18,500,1,400,30,4,65,10000,320,4500,700,36),
nrow=6, ncol=3) # setting the logical bounds for variables with missing values
> cred.aimp=amelia(cred, m=5, bound=bds, max.resample=1000) # multiple
imputation, m=5
> summary(cred.aimp) # summarizing the results
> write.amelia(cred.aimp, "C:/Documents and Settings/dane/cred_imp",
format="csv") # writing the imputed data sets to file
```

To combine the results from multiple imputation data sets the Zelig package can be used:

```
> require(Zelig)
> cred.zelig=zelig(Y~X1+X2+X3+X4+X5+X6,
data=cred.aimp$imputations, model="logit")
> summary(cred.zelig)
```

Multiple imputation performing by the Hmisc package is based on additive regression, bootstrapping and predictive mean matching techniques (the aregImpute function) or on the transformations/imputations using canonical variates (the transcan function):

```
> require(Hmisc)
> cred.Himp=aregImpute(~Y+X1+X2+X3+X4+X5+X6, n.impute=5,
data=cred)
> cred.H.fit=fit.mult.impute(Y~X1+X2+X3+X4+X5+X6, lm, cred.Himp,
data=cred) # combining the results from multiple imputation
> summary(cred.H.fit)
> cred.Himp.t=transcan(~Y+X1+X2+X3+X4+X5+X6, method="canonical",
n.impute=5, imputed=TRUE, data=cred)
> cred.H.t.fit=fit.mult.impute(Y~X1+X2+X3+X4+X5+X6, lm, cred.Himp.t,
data=cred)
> summary(cred.H.t.fit)
```

The last package used in the example is mice. Multiple imputation by chained equations method, implemented in mice, uses regression models and Bayesian sampling to impute missing values conditional on other predictors. The following list of commands should be useful to obtain the results:

```
> require(mice)
```

```
> cred.mice=mice(data=cred, m=5, seed=123) # multiple imputation by predictive mean matching
> cred.mice.fit=glm.mids(Y~X1+X2+X3+X4+X5+X6, family=binomial(link=logit), data=cred.mice) # applying glm() to a multiply imputed data set
> cred.mice.fit.pool=pool(cred.mice.fit) # pooling the results of m=5 repeated complete data analysis
> summary(cred.mice.fit.pool)
> cred.mice.sample=mice(data=cred, m=5, seed=123,imputationMethod="sample") # multiple imputation by simple random sampling
> cred.mice.sample.fit=glm.mids(Y~X1+X2+X3+X4+X5+X6, family=binomial(link=logit), data=cred.mice.sample)
> cred.mice.sample.fit.pool=pool(cred.mice.sample.fit)
> summary(cred.mice.sample.fit.pool)
```

All the results obtained from described imputation techniques are presented in Table 3. The results of classifying borrowers into the risk groups based on their predicted probabilities are summarized in Table 4.

Table 3. The results of fitting logistic regression models to imputed data sets.

Imputation method	Variable	Coeff.	SE	p-value
1	2	3	4	5
Mean Imputation	Intercept	0.74652	0.5904	0.2061
	X1	-0.01823	0.0112	0.1041
	X2	0.00043	0.0002	**0.0118**
	X3	-0.00487	0.0018	**0.0060**
	X4	-0.00056	0.0002	**0.0010**
	X5	-0.00290	0.0025	0.2478
	X6	0.01046	0.0191	0.5832
kNN Imputation (yaImpute)	Intercept	0.91901	0.5196	0.0770
	X1	-0.00773	0.0109	0.4788
	X2	0.00031	0.0001	**0.0182**
	X3	-0.00667	0.0018	**0.0002**
	X4	-0.00060	0.0002	**0.0004**
	X5	-0.00214	0.0020	0.2889
	X6	-0.00135	0.0149	0.9274

Table 3 (cont.)

1	2	3	4	5
Multiple Imputation by Bootstrapping and EM algorithm (Amelia II)	Intercept	0.34823	0.6398	0.5863
	X1	-0.01907	0.0121	0.1186
	X2	0.00020	0.0002	0.3632
	X3	-0.00446	0.0018	**0.0159**
	X4	-0.00056	0.0002	**0.0014**
	X5	0.00054	0.0033	0.8683
	X6	0.02901	0.0238	0.2228
Multiple Imputation by Additive Regression, Bootstrapping and Predictive Mean Matching techniques (Hmisc)	Intercept	0.49910	0.1459	**0.0007**
	X1	-0.00381	0.0025	0.1229
	X2	0.00002	0.0000	0.6680
	X3	-0.00111	0.0004	**0.0028**
	X4	-0.00012	0.0000	**0.0013**
	X5	0.00046	0.0007	0.5191
	X6	0.00942	0.0053	0.0751
Imputation method	Variable	Coeff.	SE	p-value
Multiple Imputation by Canonical Variates (Hmisc)	Intercept	0.51020	0.1348	**0.0002**
	X1	-0.00482	0.0024	**0.0486**
	X2	0.00000	0.0000	0.9888
	X3	-0.00103	0.0004	**0.0047**
	X4	-0.00013	0.0000	**0.0004**
	X5	0.00081	0.0006	0.2067
	X6	0.01076	0.0050	**0.0312**
Multiple Imputation by Chained Equations using Predictive Mean Matching (mice)	Intercept	0.51010	0.1473	**0.0006**
	X1	-0.00340	0.0026	0.1978
	X2	0.00002	0.0000	0.6038
	X3	-0.00114	0.0004	**0.0071**
	X4	-0.00012	0.0000	**0.0013**
	X5	0.00039	0.0007	0.5858
	X6	0.00868	0.0051	0.0916

Table 3 (cont.)

1	2	3	4	5
Multiple Imputation by Chained Equations using Simple Random Sampling (mice)	Intercept	0.62004	0.1255	**0.0000**
	X1	-0.00396	0.0025	0.1075
	X2	0.00007	0.0000	**0.0295**
	X3	-0.00110	0.0004	**0.0042**
	X4	-0.00012	0.0000	**0.0009**
	X5	-0.00030	0.0005	0.5783
	X6	0.00406	0.0037	0.2741

Source: Author's calculations.

Table 4. Proportions of correctly classified objects.

Method	% of correct classifications
Original data set (no missing values)	64.88%
Complete Case Analysis	62.24%
Mean Imputation	64.03%
kNN Imputation (yaImpute)	65.52%
Multiple Imputation by Bootstrapping and EM algorithm (Amelia II)	64.24%
Multiple Imputation by Additive Regression, Bootstrapping and Predictive Mean Matching techniques (Hmisc)	64.03%
Multiple Imputation by Canonical Variates (Hmisc)	63.81%
Multiple Imputation by Chained Equations using Predictive Mean Matching (mice)	64.88%
Multiple Imputation by Chained Equations using Simple Random Sampling (mice)	63.38%

Source: Author's calculations.

Since an example is presented (not a simulation study), there is no possibility to draw general conclusions but, as we can see, the worst results are obtained from complete case analysis. Only one coefficient in logistic regression model is significant and the misclassification error rate is the highest. Imputation procedures lead to quite similar results concerning both the logistic regression model and the misclassification error rate.

V. CONCLUDING REMARKS

The objective of the paper was to review selected imputation techniques. Special attention was paid to methods implemented in some packages working in the R environment. The goal of the example was only to show how to handle missing values using a few procedures implemented in R and not to compare any imputation techniques.

Ambler et al. (2007) presented the results of a simulation comparison of different imputation techniques for handling missing predictor values in a risk model based on logistic regression. They showed that missing data could affect the predictions from risk models and simply ignoring missing data and performing a complete case analysis could lead to substantial bias and poor predictions. Single imputation procedures improved the results but they did not allow for imputation uncertainty so the confidence intervals of the regression coefficients could be too narrow and p-values too small. The best way to handle missing data is multiple imputation. Multiple imputation techniques generally performed well and they should be recommended in practical applications.

REFERENCES

Allison P. D. (2002), *Missing data*, Series: Quantitative Applications in the Social Sciences 07-136, SAGE Publications, Thousand Oaks, London, New Delhi.

Ambler G., Omar R. Z., Royston P. (2007), *A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome*, "Statistical Methods in Medical Research" 2007; 16: 277–298.

Crookston N. L., Finley A. O. (2008), *yaImpute: An R Package for kNN Imputation*, "Journal of Statistical Software", January 2008, Volume 23, Issue 10.

Horton N. J., Kleinman K. P. (2007), *Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models*, "The American Statistician" 2007, 6 (1): 79-90.

Kenward M. G., Carpenter J. (2007), *Multiple imputation: current perspectives*, "Statistical Methods in Medical Research" 2007; 16: 199–218.

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

Molenberghs G., Kenward M. G (2007), *Missing Data in Clinical Studies*, Wiley, England.

Schafer J. L. (1996), *Analysis of Incomplete Multivariate Data,* Chapman & Hall, New York.

Su Y.-S., Gelman A., Hill J., Yajima M. (2011), *Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box*, "Journal of Statistical Software", in press.

van Buuren S., Groothuis-Oudshoorn K. (2011), *MICE: Multivariate Imputation by Chained Equations in R*, „Journal of Statistical Software", in press.

Wayman J. C. (2003), *Multiple Imputation for Missing Data: What Is It And How Can I Use It?*, http://www.csos.jhu.edu/contact/staff/jwayman_pub/ wayman_multimp_aera2003.pdf.

Yu L.-M., Burton A., Rivero-Arias O. (2007), *Evaluation of software for multiple imputation of semi-continuous data*, "Statistical Methods in Medical Research" 2007; 16: 243–258.

Małgorzata Misztal

IMPUTACJA BRAKUJĄCYCH DANYCH Z WYKORZYSTANIEM ŚRODOWISKA R

W praktycznych zastosowaniach metod statystycznych często pojawia się problem występowania w zbiorach danych brakujących wartości. W takich sytuacjach wykorzystać można metody imputacji danych, polegające na zastąpieniu brakujących danych konkretnymi wartościami w celu uzyskania kompletnego zbioru danych.

W referacie dokonano przeglądu metod imputacji danych oraz opisano możliwości wykonania koniecznych obliczeń z wykorzystaniem dostępnych w środowisku R pakietów realizujących procedury imputacji jednostkowej i wielokrotnej.