REVIEW OF METHODS FOR DATA SETS WITH MISSING VALUES AND PRACTICAL APPLICATIONS

ŚLĄSKI
PRZEGLĄD
STATYSTYCZNY
Nr 12(18)

Adam Korczyński
Warsaw School of Economics

ISSN 1644-6739

Summary: The aim of this paper is to revise the traditional methods (*complete-case analysis*, *available-case analysis, single imputation*) and current methods (*likelihood-based methods, multiple imputation, weighting methods*) for handling the problem of missing data and to assess their usefulness in statistical research. The paper provides the terminology and the description of traditional and current methods and algorithms used in the analysis of incomplete data sets. The methods are assessed in terms of the statistical properties of their estimators. An example is provided for the multiple imputation method. The review indicates that current methods outweigh traditional ones in terms of bias reduction, precision and efficiency of the estimation.

Keywords: missing data pattern, missing data mechanism, complete-case analysis, available-case analysis, single imputation, likelihood-based methods, multiple imputation, weighting methods.

DOI: 10.15611/sps.2014.12.05

1. Missing data – fundamental concepts

Most statistical methods assume that the characteristics of subjects examined are fully observed. In the case of a survey this means that every respondent has divulged the information needed to complete the whole questionnaire. In practice this happens very rarely, if ever. Usually, the researcher is confronted with a *unit nonresponse* as well as an *item nonresponse* [Balicki 2004]. Thus one can ask how to eliminate or at least reduce bias caused by missing data? The answer to this question requires an insight into the nature of missingness and an in-depth analysis of its hypothetical influence on the results of the research.

1.1. Objectives of the analysis

The features of a good method for handling the problem of missing data are the following [Rubin 1987, p. 11; Allison 2009, p. 75; Graham 2012, p. 5]:

1) minimizing the bias of the parameter estimates,

2) using the maximum amount of information from the data set (increase of efficiency),

3) producing appropriate estimates of uncertainty (standard errors and confidence intervals).

Substantial number of methods for incomplete data sets are based on imputation [Little, Rubin 2002, pp. 59–90]. However, is it worth emphasizing that imputation itself is not the objective of the analysis. In other words, the idea is not about the prediction of missing values but about estimation (properties of the estimators) and its sensitivity versus the missingness; still, it is advised to check if the imputed values are from the specified range. Graphical analysis of the distribution (histogram or box-and-whisker plot) of imputed values can be applied for that purpose [White 2013, p. 18].

1.2. Mechanism leading to missingness

The analysis of the incomplete data set requires an insight into the nature of the process which leads to missingness. The basic missingness mechanisms distinguished in the statistical literature are the following (see [Little, Rubin 2002, pp. 11–13]):

– *missing completely at random* (MCAR),
– *missing at random* (MAR),
– *not missing at random* (NMAR).

We may say that data are MCAR if missingness does not depend on both observed and unobserved outcomes. In this case, the observed outcomes are a random subsample of the original sample and thus one can estimate the parameter of interest without bias. For instance, the data are MCAR if the interviewer conducting a budget survey fails to reach the respondent due to illness.

The missing completely at random assumption is the most restrictive one. If it is rejected, one can assume two mechanisms will remain valid, namely MAR or NMAR.

Under MAR assumption, the missingness depends on the observed outcomes and not on the unobserved part of the outcome. For example, missingness of data on the use of social services possibly depends on the social status of a household. The missing at random assumption corresponds to the statement that the unobserved data and the observed data are related and one can correct for the missingness given the observed values. Valid and efficient likelihood-based or Bayesian inference about a given parameter requires that the data are MAR and a so called separability condition is satisfied [Little, Rubin 2002, pp. 118––122]. If this is the case, we say that the missingness is *ignorable*. In practice it means that "we can construct a valid analysis that does not

require us to explicitly include the model for that missing value mechanism" [Carpenter, Kenward 2013, p. 34]. MAR assumption is not subject to testing. However, we can make it more plausible by putting many variables in the imputation model [Allison 2009, p. 74].

In practical settings one can arrive at the case when the missingness depends on the unobserved outcomes. In that situation the values are said to be *not missing at random* (NMAR). In order to imagine a situation when the NMAR assumption might be plausible we can refer to the income questions in budget surveys. These can be considered as NMAR because the level of income might possibly influence the propensity of its reporting. When the MAR assumption is violated and the data are said to be NMAR, the missingness mechanism has to be modelled [Allison 2009, p. 74]. Modelling of the missingness mechanism requires a deep prior knowledge about the nature of missingness. In that case the researcher should conduct a sensitivity analysis thus allowing an assessment of the sensitivity of the estimates to the assumptions concerning unobserved data.

2. Methods for handling missingness

When trying to build a taxonomy of missing-data methods we can look at them from at least two perspectives: over time and over their type. Little and Rubin [2002, pp. 19-20] have grouped the missing-data methods into the following categories:

– *Procedures based on completely recorded units*.
– *Weighting procedures*, the methods which adjust the design weights for a nonresponse. The adjustment is based on an estimate of the probability of response conditioned by the selection to the sample according to the following product rule property of probability of two random events [Little, Rubin 2002, p. 46]:
  Pr(selection and response)=Pr(selection) · Pr(response|selection).
  For the estimate of the completers' fraction we can take the proportion of responding units in a specified subclass.
– *Imputation-based procedures*, in which the missing values are filled in by a constant (e.g. mean) or by a value observed in a similar object (e.g. hot deck, cold deck).
– *Model-based procedures* referring to the likelihood or posterior distribution (Bayesian approach) of the observed data. The method is applied especially in the case of NMAR data.

Since the problem of missing data is not new in statistical analysis,[1] it has recently become popular to speak about traditional and contemporary methods for handling missingness. Both types of methods are presented and discussed in [Little, Rubin 2002; Molenberghs, Kenward 2007; Allison 2009].

2.1. The traditional methods for handling nonresponse in surveys

The following subsection contains a description of traditional methods used for data sets with missing values. The list of methods commences with the simplest ones (*complete-case analysis* and *available-case analysis*) in which the missing values are simply discarded from the analysis. The remaining methods (*dummy variable adjustment*, *unconditional mean imputation*, *regression imputation*, *stochastic regression imputation*, *hot* and *cold* deck *imputation*) are based on *single imputation*.

Complete-case analysis is also known as *casewise deletion* and *listwise analysis*. This is the simplest method in which we ignore the observations with missing data. It is valid only under MCAR. *Complete-case analysis* leads to a significant loss of efficiency of the estimators if the fraction of missing data is large.

Available-case analysis (*pairwise deletion*) includes all cases for which the variable or variables of interest are observed. In the case of bivariate analysis with r_1 observed outcomes for the first variable, and r_2 outcomes observed for the second one, the estimates of means are based on r_1 and r_2 observations respectively. This method recovers part of the information lost in *complete-case analysis*, however, it yields practical and theoretical problems related to the variability of the sample base. It is valid under MCAR.

Dummy variable adjustment is a method used in the regression analysis. Let us assume we have an outcome variable Y and some predictors X_j ($j = 1, 2, \ldots, r$). For each variable with missing data we are creating a dummy missingness indicator variable M_j. The regression model includes both X and M. The missing values of X are replaced by a constant, for instance, by the mean for non-missing values. The method enables use of all the information from the data set, however, it produces biased estimates of the regression coefficients, even under MCAR.

[1] Ronald Fisher in his work *The Design of Experiments* [1971, p. 177–180] first published in 1935 has suggested the single imputation method of correction for the problem of missing values in analysis of variance. Van Buuren [2012, p. 25] has mentioned Allan and Wishart [1930] as those who were the first to develop a statistical method to replace a missing value.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 12 (18)

A broad class of methods for handling nonresponse in surveys is based on *single imputation*. The basic *single imputation* methods are the following:

Unconditional mean imputation consists of filling in a constant value for each missing value. This method leads to an underestimation of the variance and thus to confidence intervals that are too short. Under MCAR one can correct for the underestimation of variance (see [Rubin 1987, pp. 13–15]). However, the method distorts the empirical distribution of the variable of interest and does not produce valid estimates of covariance.

Regression imputation, in which the missing values are replaced by the values predicted by a regression model of the variable with missingness on the fully observed variables. The method is problematic because it underestimates the standard errors of the estimates and thus leads to false (too short) confidence intervals of the parameter estimates. Moreover, in practice, it is limited to *complete-case analysis* of predictors. The method leads to consistent estimates of means under MCAR and MAR, however, additional assumptions are required in this case [Little, Rubin 2002, pp. 63–64].

Stochastic regression imputation is a development of *regression imputation*. The method consists of filling-in missing values with values predicted by a regression model corrected by a random deviation. The idea behind the method is to eliminate the problem of underestimation of uncertainty about the parameter estimates. In the linear regression models the residual values are drawn from a normal distribution with a zero mean and variances computed for the complete cases. It can be shown that in the bivariate model with one fully observed variable Y and a second one with MAR data X, the *stochastic regression imputation* leads to consistent estimates of mean, variance of X, and regression coefficients of Y on X and X on Y (see [Little, Rubin 2002, pp. 65–66]). The main drawback of the method is that it does not incorporate the imputation uncertainty which results in too small standard errors of the estimates.

The broad class of imputation methods uses the information about similarity between the objects under analysis. The two main methods in that class are the following: *single hot deck imputation*, *single cold deck imputation*.

Single hot deck imputation replaces the missing values by values observed in similar objects from the same experiment. The *single hot deck imputation* is widely used in surveys in a practical way.

Single cold deck imputation, in which the missing values are filled in by values coming from external sources, for example, from previous panels or similar objects from other studies. Since *cold deck imputation* explicitly assumes that some characteristics are stable over time and/or space it is a questionable approach.

The main advantage of traditional methods is their simplicity. However, their theoretical properties are fairly unsatisfactory mainly because they are discarding the uncertainty about the missing data and in general they are followed by standard analysis. In results the estimates obtained by traditional methods are not statistically valid or at the very least not efficient. They use is limited to rare situation when data are missing completely at random and the number of missing values is trivial. In these circumstances the sample saves its representativeness and the parameters of interest can be estimated without bias.

2.2. The current methods for handling nonresponse in surveys

The modern missing-data methods have emerged in response to the deficiencies of traditional approaches. Their development was conditioned by the significant increase in computing power dating from the early 1980s (the beginning of the microcomputer era) to the present day.

We can classify the modern methods into three main groups:
– likelihood-based methods,
– multiple imputation, and
– weighting methods.

2.2.1. Likelihood-based methods

Likelihood-based methods are concerned with inference based on the observed data likelihood modelled under specified assumptions. The parameters are estimated using maximum likelihood or Bayesian estimation. The main drawback of these types of methods is that they make the analysis complicated and result in the use of sophisticated missing-data models which are beyond the scope of the researcher's interests. However, by using them one becomes more aware in terms of the possibility of modelling the missing data mechanism by some explicit assumptions which help to understand the influence of non-response on the results.

In the likelihood-based approach we can distinguish between the *direct likelihood method* and the *iterative method* involving the expectation maximization (EM) algorithm.

2.2.1.1. Direct likelihood method

Direct likelihood method can be applied to models estimated by maximum likelihood [Little, Rubin 2002, pp. 97–112]. Basically, the idea is to use the available cases but correcting the estimates for the uncertainty coming from missing values. The implicit assumption is that the data are MAR on given variables. The application of direct likelihood method is limited because of its complexity especially in the case of non-Gaussian data.

Let us consider a bivariate case. Suppose we have a sample of n observations on two variables (Y_1, Y_2) from bivariate normal distribution. Y_1 is completely observed and Y_2 is observed only for r units. The unobserved units $\{i: r+1, r+2, \ldots, n\}$ are MAR. Our aim is to estimate μ_2, the mean of Y_2. Little and Rubin [2002, pp. 135–143] have derived a formula for a regression maximum likelihood estimator $\hat{\mu}_2$ of μ_2:

$$\hat{\mu}_2 = \overline{y}_2 - \frac{s_{12}}{s_{11}}(\overline{y}_1 - \hat{\mu}_1), \tag{1}$$

where:

$$\overline{y}_j = \frac{1}{r}\sum_{i=1}^{r} y_{ij}, \tag{2}$$

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} y_{i1}, \tag{3}$$

$$s_{jk} = \frac{1}{r}\sum_{i=1}^{r}(y_{ij} - \overline{y}_j)(y_{ik} - \overline{y}_k). \tag{4}$$

In practice the method is based on standard *regression imputation*. However, unlike the *regression imputation* the *direct likelihood method* leads to unbiased estimates of standard errors of the estimates, which results in valid confidence interval estimates. The large sample $(1 - \alpha)\cdot 100\%$ confidence interval for $\hat{\mu}_2$ is given by:

$$\hat{\mu}_2 \pm u_\alpha \sqrt{D^2(\hat{\mu}_2 - \mu_2)}, \tag{5}$$

where u_α is a critical value form standard normal distribution. The variance of the distance between $\hat{\mu}_2$ and μ_2 is approximated by:

$$D^2(\hat{\mu}_2 - \mu_2) = \left(s_{22} - \frac{s_{12}^2}{s_{11}}\right)\left(\frac{1}{r} + \frac{\hat{\rho}}{n(1-\hat{\rho}^2)} + \frac{(\overline{y}_1 - \hat{\mu}_1)^2}{r s_{11}}\right), \tag{6}$$

where:

$$\hat{\rho} = \frac{s_{12}}{\sqrt{s_{11}s_{22}}} \sqrt{\frac{\hat{\sigma}_{11}}{s_{11}}} \sqrt{\frac{s_{22}}{\hat{\sigma}_{22}}}, \tag{7}$$

and:

$$\hat{\sigma}_{11} = \frac{1}{n} \sum_{i=1}^{n} (y_{i1} - \hat{\mu}_1)^2, \tag{8}$$

$$\hat{\sigma}_{22} = s_{22} - \frac{s_{12}^2}{s_{11}} + \left(\frac{s_{12}}{s_{11}} \right)^2 \hat{\sigma}_{11}. \tag{9}$$

The direct likelihood method described above can be applied to a multivariate case with a monotone missing data pattern [Little, Rubin 2002, pp. 143–155]. The missing data pattern is said to be monotone if the set of variables $\{Y_j: j = 1, 2, \ldots, k\}$ can be ordered in such a way that when a missing value appears for an observation on variable Y_j all of the consecutive variables Y_{j+1}, Y_{j+2},…, Y_k have a missing value for that observation. The monotonicity of missingness is required because, as mentioned above, in practice the *direct likelihood method* uses regression imputation, which is recursive in this case.

The main drawback of the method is that it assumes that at least one variable is fully observed and that the missing data pattern is monotone. For the missing-data general patterns the EM algorithm seems an attractive computational alternative [Little, Rubin 2002, pp. 164-188].

2.2.1.2. The EM algorithm

The EM algorithm is an iterative counterpart of the direct likelihood method. The *EM algorithm* enables us to identify the maximum likelihood estimates (see [Little, Rubin 2002; Schafer 1997; Molenberghs, Kenward 2007; Zdobylak, Zmyślona 2004]). It is less computational cumbersome than the direct likelihood method. The algorithm is based on two major steps: the expectation step (E) in which we are calculating the expected value of the log-likelihood for a given data and the maximization step (M) in which the log-likelihood is maximized. The two steps are repeated till the convergence of the algorithm is reached. The main advantage of the *EM algorithm*, apart from its computational simplicity is that it can be used for general patterns of missing data. As regards the main drawbacks of the method Little and Rubin [Little, Rubin 2002, p. 166] mention problems with reaching convergence in data sets with a large fraction of missing values

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 12 (18)

and the difficulties with the M step. Furthermore, the *EM algorithm* does not yield the standard errors of the estimates, thus, additional computations are required to produce estimates of the precision of the estimation (see [Molenberghs, Kenward 2007, p. 98]).

Let us depict the steps of the algorithm in its general form. Suppose we have a data set with n observations and k variables. Let $\boldsymbol{Y}_{nxk} = [y_{ij}]$ ($i = 1, 2, \ldots, n; j = 1, 2, \ldots, k$) denote the complete data; assuming a presence of missing values, complete data set can be divided into an observed and unobserved part $\boldsymbol{Y}_{nxk} = (\boldsymbol{Y}^{\text{obs}}, \boldsymbol{Y}^{\text{mis}})$. For example, for jth variable with r missing values we have $\{y_{ij}^{\text{obs}}: i = 1, 2, \ldots, r\}$ and $\{y_{ij}^{\text{mis}}: i = r + 1, r + 2, \ldots, n\}$. Having all variables divided into observed and missing part one can determine the missing data patterns as it is shown in Figure 1. Each iteration of the *EM algorithm* consists on computations carried out for each pattern separately. In practice, for the case of data set given in Figure 1 the expected values of Y_{ij}^{mis}, $(Y_{ij}^{\text{mis}})^2$ and $Y_{ij}^{\text{mis}} Y_{il}$ (for $j \neq l$) conditional on covariates are calculated for each pattern. These expected values are then add up and used to update the parameter estimates.

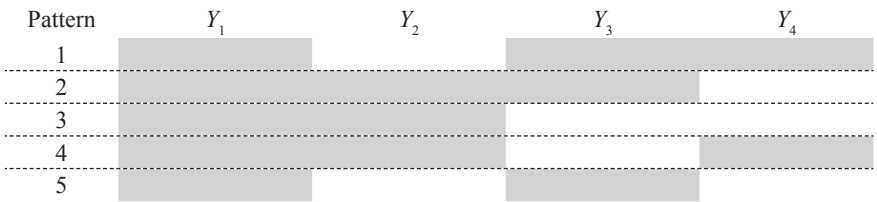

Figure 1. Example of a missing data pattern. Grey area stands for observed values, white stands for missing ones

Source: own elaboration.

In order to describe the *EM algorithm* let us assume that the aim of the analysis is to find the ML estimates of the parameter vector $\boldsymbol{\theta}$. For instance it could be a vector of means $\boldsymbol{\mu} = [\mu_1\ \mu_2]^T$. The *EM algorithm* is based on the following steps:

Step 1. Set the initial value of the $\boldsymbol{\theta}^{(0)}$. The value can by calculated using a traditional method like *complete-case analysis* or *available-case analysis.*

Step 2. (E step) Calculate the expected value of the log-likelihood $\ln L(\boldsymbol{\theta} \mid \mathbf{Y})$ given the observed data and the current value of $\boldsymbol{\theta}^{(t)}$ where t denotes the iteration number ($t = 0, 1, 2, \ldots, l$). For the first iteration we set $\boldsymbol{\theta}^{(t)}=\boldsymbol{\theta}^{(0)}$. For the t-th iteration we have

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = E\left(\ln L(\boldsymbol{\theta} \mid \mathbf{Y}) \mid \mathbf{Y}^{obs}, \boldsymbol{\theta}^{(t)}\right) \tag{10}$$

Step 3. (M step) Calculate $\boldsymbol{\theta}^{(t+1)}$ the parameter vector that maximizes the expected log-likelihood $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ from the E step. Note that in this step we are maximizing the log-likelihood of imputed data set. As a result we get:

$$Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}), \text{ for all } \boldsymbol{\theta}. \tag{11}$$

With $\boldsymbol{\theta}^{(t+1)}$ go to step 2. The cycle repeats until the convergence of the algorithm.

As claimed before, the likelihood-based methods do not provide a straightforward way to calculate the standard errors of the estimates. In order to assess the precision of the estimation we can refer to *multiple imputation*.

2.2.2. Multiple imputation

In the *multiple imputation method* we replace the missing values by several values, say *m* [Zdobylak, Zmyślona 2004; Zmyślona 2011]. Each data set is then analyzed by standard complete-data method, yielding *m* estimates. The estimates are combined in order to make a single inference. The uncertainty has two separate sources: the sampling variability and variability caused by lack of sufficient knowledge about the actual reasons for nonresponse. The sampling variability is incorporated by the use of several imputations under one imputation model. The nonresponse variability is expressed by the use of several imputation models. The main drawback of the method is in its complexity and sensitivity to the choice of imputation models. Moreover, it is not straightforward if the methods are applied to discrete data [Carpenter, Kenward, Vansteelandt 2006, pp. 12–13].

The multiple imputation method is comprised of the two major steps:

Step 1. Define the imputation model. The two basic groups of imputation models considered in statistical literature are the regression models and models based on the *hot deck* procedure. In the *hot deck* method units of observation are divided in separate groups with different probability of missingness and then the missing values are replaced by draws from the observed values made within homogenous groups (see for example [Allison 2000; Heitjan, Little 1991]).

Step 2. Analyze the imputed data sets using standard a complete-data method (e.g. analysis of variance or linear regression). Let *m* denote the number of imputations. Each imputation yields an estimate

$\boldsymbol{\theta}_d$ (d = 1, 2, …, m) of the parameter of interest $\boldsymbol{\theta}$ and its covariance matrix $\boldsymbol{S}_d$. The combined estimate of $\boldsymbol{\theta}$ is defined as the average over the $\boldsymbol{\theta}_d$ estimates:

$$\hat{\boldsymbol{\theta}} = \frac{1}{m}\sum_{d=1}^{m}\boldsymbol{\theta}_d \,. \tag{12}$$

The estimate of the covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by:

$$\mathbf{T} = \mathbf{W} + \left(\frac{m+1}{m}\right)\mathbf{B}\,, \tag{13}$$

where:

$$\mathbf{W} = \frac{1}{m}\sum_{d=1}^{m}\mathbf{S}_d\,, \tag{14}$$

$$\mathbf{B} = \frac{1}{m-1}\sum_{d=1}^{m}\left(\boldsymbol{\theta}_d - \hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta}_d - \hat{\boldsymbol{\theta}}\right)^T. \tag{15}$$

Component $\boldsymbol{W}$ of the formula (13) reflects the within-imputation variability, say the sampling variability. The $\boldsymbol{B}$ component stands for the between-imputation variance, a measure of the variability between the estimates from different imputations. The test and confidence intervals for *multiple imputation* estimates are based on [Molenberghs, Kenward 2007, pp. 108–109]:

$$F_{k,w} = \frac{\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^T \mathbf{T}^{-1}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)}{k(1+\gamma)}\,, \tag{16}$$

where k is the number of estimated parameters, $\boldsymbol{\theta}_0$ is the vector of parameters under null hypothesis and:

$$\gamma = \frac{1}{k}\left(1+\frac{1}{m}\right)tr\left(\mathbf{B}\mathbf{W}^{-1}\right), \tag{17}$$

$$w = 4 + \left(k\gamma - k - 4\right)\left(1 + \frac{1-2\left[k\gamma - k\right]^{-1}}{\gamma}\right)^2. \tag{18}$$

For a scalar parameter θ, assuming large sample sizes, the $(1-\alpha)\cdot 100\%$ interval estimate is [Little, Rubin 2002, p. 87]:

$$\hat{\theta} \pm t_{\alpha;v}\sqrt{T}\,, \tag{19}$$

where t denotes the t distribution with v degrees of freedom given by:

$$v = (m-1)\left(1+\frac{W}{(1+m^{-1})B}\right)^2. \tag{20}$$

The main advantage of the multiple imputation method is that it leads to statistically valid estimates under MAR assumption. The method is relatively simple and easy to apply in practice. It enables us to control the imputation model and to incorporate the uncertainty about the imputed values. The remarkable advantage of multiple imputation is that it can be implemented for general patterns of missing data and it allows for the sensitivity analysis under different imputation models which is critical when dealing with NMAR data.

2.2.3. Weighting methods

The weighting methods were developed for sampling schemes in which each unit selected to the sample represents a specified number of units in the population. There are two basic classes of *weighting methods*:

– *calibration* (*weighting in sampling schemes*) which can be viewed as a development of standard sampling methods in which one takes into account the fact that units are observed only with a certain probability (see [Paradysz, Szymkowiak 2007; Szymkowiak 2012]),
– *inverse probability weighted estimating equations* – a semiparametric method of correction for missingness in the estimation of linear models [Molenberghs, Kenward 2007].

2.2.3.1. Weighting in sampling schemes

Let π_i where $i = 1, 2, ..., n$ denotes the selection probability and φ_i denotes the probability of response conditioned by the selection to the sample for unit i. The probability of selection and response for unit i is equal to $\pi_i\varphi_i$. Let us assume that only r units of the initial sample of size n are observed. The Horvitz-Thompson estimator of mean of Y is given by [Bracha 1998, pp. 195–198]:

$$\overline{y}_o = \frac{\sum_{i=1}^{r}\frac{Y_i}{\pi_i\varphi_i}}{\sum_{i=1}^{r}\frac{1}{\pi_i\varphi_i}}, \tag{21}$$

Suppose the sample units can be divided into groups of equal probability of response. The probability of response for sample unit i in group j (j = 1, 2, ..., k) can be estimated by:

$$\hat{\varphi}_i = \frac{r_j}{n_j}, \tag{22}$$

where r_j denotes the number of observed units in group j and n_j denotes the size of group j. For a simple random sample without replacement the selection probability $\pi_i = \frac{n}{N}$ is the same for each unit. Assuming that the response probabilities of sampled units φ_i are correctly specified then the unbiased estimator of the mean of Y is given by:

$$\overline{y}_o = \frac{1}{n}\sum_{j=1}^{k} n_j \overline{y}_{oj}\,, \tag{23}$$

where

$$\overline{y}_{oj} = \frac{1}{r_j}\sum_{i=1}^{r_j} y_{ij} \tag{24}$$

is the sample mean for the j group. If Y is *missing at random* (MAR), the estimator (23) is unbiased and its variance and mean square error are equal. The variance of (23) is given by [Little, Rubin 2002, p. 47]:

$$D^2(\overline{y}_o) = \sum_{j=1}^{k}\left(\frac{n_j}{n}\right)^2\left(1-\frac{r_j n}{n_j N}\right)\frac{s_{oj}^2}{r_j} + \frac{N-n}{(N-1)n^2}\sum_{j=1}^{k} n_j\left(\overline{y}_{oj}-\overline{y}_o\right)^2, \tag{25}$$

where

$$s_{oj}^2 = \frac{1}{r_j}\sum_{i=1}^{r_j}(y_{ij}-\overline{y}_{oj})^2\,.$$

The $(1-\alpha)\cdot 100\%$ confidence interval for mean of Y can be then constructed by:

$$\overline{y}_o \pm u_\alpha D(\overline{y}_o)\,,$$

where u_α is the $100\,(1-\alpha/2)$ percentile of standard normal distribution.

2.2.3.2. Inverse probability weighted estimating equations

Suppose our aim is to regress Y on the covariates X_1, X_2, ..., X_k. The vector of observations of unit i is $\mathbf{x}_i = [1\ x_{i1}\ \ldots\ x_{ik}]^T$ and the vector of parameters is $\boldsymbol{\beta} = [\beta_0\ \beta_1\ \ldots\ \beta_k]^T$. The linear regression model to be estimated is given by:

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i \,. \tag{26}$$

Further, let us assume that Y is observed for r out of n sampled units and the covariates are fully observed. Define an indicator $n \times 1$ vector M with elements $m_i = 0$ when Y is observed and $m_i = 0$ is missing. Moreover, let us assume that Y is MAR on covariates. The response probability[2] for sampled unit i is [Little, Rubin 2002, pp. 49–50]:

$$\varphi_i = P(m_i = 0 \,|\, x_1, x_2, ..., x_k) \,. \tag{27}$$

Given the response probabilities one can correct for missingness by plugging them into normal equations:

$$\sum_{i=1}^{r} \frac{1}{\varphi_i} \left(\frac{\partial \mathbf{x}_i^T \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} \right)^T \left(Y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \right) = \mathbf{0} \,. \tag{28}$$

The solution of (27) in respect of $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}_1 & ... & \hat{\beta}_k \end{bmatrix}^T$ gives the *least squares* estimators of regression coefficients of (26). If the response probability weights are correctly specified i.e. the model estimating φ_i takes into account all variables and interactions which influence the response probability, the estimates of $\boldsymbol{\beta}$ are consistent.

The *weighting methods* are applied especially when it is necessary to take into account the sampling scheme. The disadvantage of this group of methods is that they require the proper specification of response probabilities φ_i. Moreover, the efficiency of the *weighting methods* estimates is sensitive to the response rate. Namely, with the decrease of responding units, the efficiency of the estimates decrease because the estimate is based only on the observed units. In more complex settings the calculation of appropriate standard errors requires computationally intensive approaches [Little, Rubin 2002, p. 53].

3. Example

In the example the *multiple imputation* method as well as the *complete-case* and *available-case* methods were used to study the level and change of the real disposable income (Y_t)[3] of adult members of Polish households in 2007, 2009 and 2011.

---

[2] The response probabilities can be estimated by logistic regression model of M on X.

[3] Data set is from the panel survey "Diagnoza Społeczna" [Social Diagnosis 2000–2013]. The nominal income was adjusted according to the Consumer Price Index published by GUS [2013].

The aim of this exemplification is to compare the estimates obtained with a use of traditional methods (a *complete-case* and *available-case* methods) and a *multiple imputation* method. The application of a *multiple imputation* method will allow us to assess the potential bias of a traditional method under the plausible assumption of Y_t missing-at-random.

The analysed data set contains n = 5415 observations[4] on the following variables:

– *real disposable income in 2007, 2009 and 2011*,
– *age of household member in 2011*,
– *years spent in education of the household member in 2011*,
– *number of people in the household in 2007, 2009 and 2011*,
– *working status of the household member in 2007, 2009 and 2011* (1 = yes, 0 = no).

Let us consider a panel study with Y_t (t = 1, 2, ..., l) subject to nonresponse. We assume that missingness occurs only for an *income* variable and the predictors used in the imputation model are fully observed. Empirical observation shows that income value is likely to be missing. In our example, in each wave almost 16% of respondents have their income value missing whereas the second least frequently observed variable is *working status* with less than 1% of missing data for each wave. In total 154 observations (2.8% of the sample size) were discarded from the study so as to make the assumption about missingness on *income* valid. The argument for this approach is that it leads to a relatively small loss of information and makes the problem more tractable. The patterns of missingness in the final data set are given in Table 1.

The analysed data set contains n = 5415 observations. Of these, 3700 individuals reported the value of *income* in each wave of survey, and 1715 did not (Table 1).

Figures 2 and 3 present the distribution of income respectively for the *complete-case* (CC) and *available-case* (AC). As we can see, a *complete-case* and *available-case* analysis of income distribution show that the average real disposable income of members of Polish households increased comparatively in 2007, 2009 and 2011. Figures 2 and 3 indicate that the income distribution is more platykurtic in 2009 and 2011 than at the beginning of the study in 2007. The level of missingness

[4] The original data set is limited to the household members being at least 18 years old in 2007 with known working status in the consecutive years of survey (2007, 2009, 2011).

Table 1. Pattern of missing data for *income* in the consecutive years of the survey (grey = observed, white = missing)

Pattern	Income in 2007	Income in 2009	Income in 2011	*nr**	*rr***
1	■	■	■	3700	68%
2	■	■		355	7%
3	■		■	324	6%
4	■			186	3%
5		■	■	400	7%
6		■		101	2%
7			■	160	3%
8				189	3%
Total	4565	4556	4584	5415	100%

**nr* – number of observations with reported value of income;
***rr* – percent of observations with reported value of income.

Source: own elaboration based on Social Diagnosis data [2000–2013].

of income values was 31.7% for *complete-cases* analysis and approximately 15.5% in each year for *available-case* analysis. Thus we can expect that in each case the results are biased.

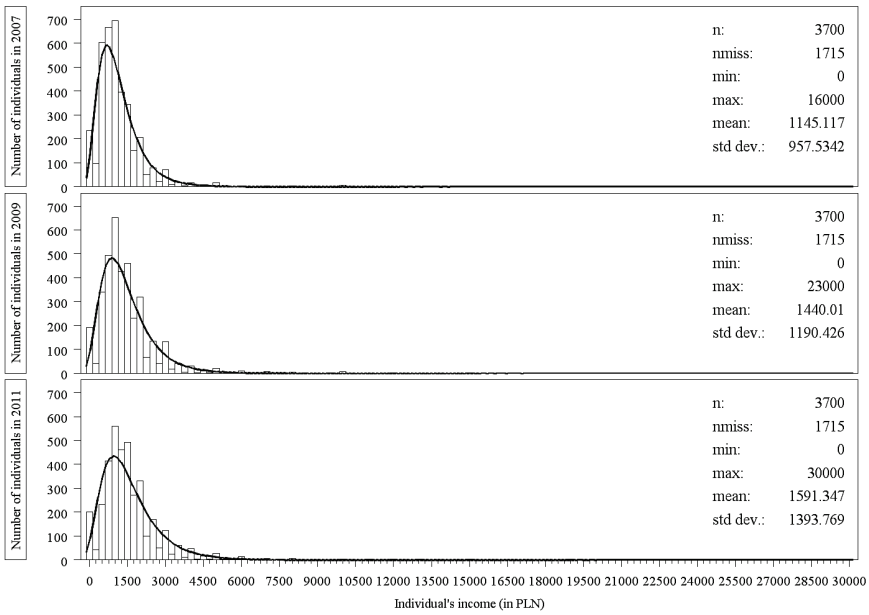

Figure 2. Distribution of income of surveyed household members – *complete-case* (CC) analysis

Source: own elaboration.

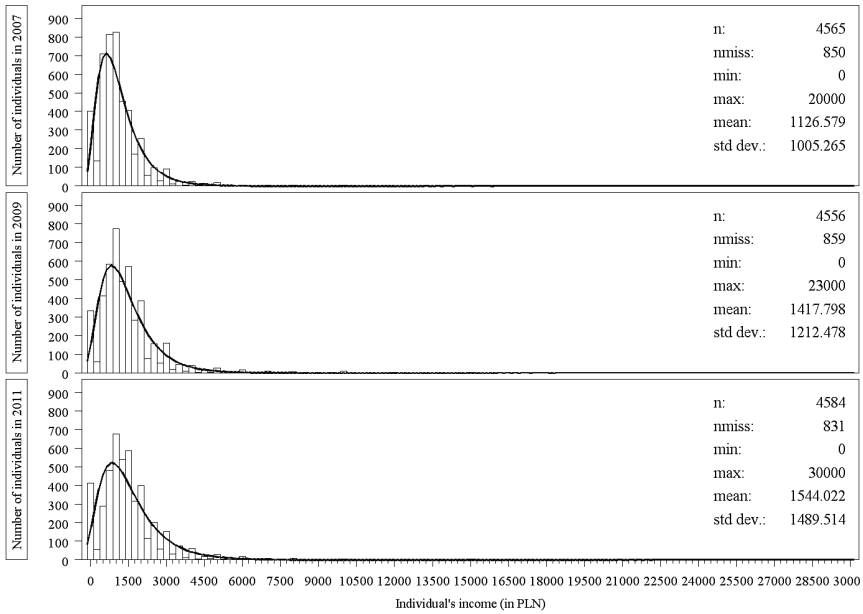

Figure 3. Distribution of income of surveyed household members – *available-case* (AC) analysis

Source: own elaboration.

In our example the substantive parameter is the average real disposable income for any given moment μ_t^Y [Tian 2005]:

$$\mu_t^Y = (1-\delta_t)\exp(\mu_t + \sigma_t^2/2). \tag{29}$$

where δ_t is the fraction of households having zero incomes. The distribution of non-zero incomes is said to be lognormal with parameters μ_t and σ_t^2. The choice of μ_t^Y is driven by the nature of households income distribution which has a lognormal shape except for those respondents reporting no income[5]. The factor $(1 - \delta_t)$ captures the probability of having positive income. Thus μ_t^Y given by (29) can be perceived as conditional expected value of income given that it is greater than zero.

The imputation model applied for that data is based on the fully conditional specification (FCS) in which the variables of interest are

[5] In the econometric literature income distribution is commonly approximated by the lognormal probability density function (see for example [Kalecki 1945]). However, it is often the case that a part of the respondents report zero income for a particular moment of observation. The estimate of the average μ_t^Y of the random variable Y with set of zeros and lognormal distribution for $Y > 0$ can be found in [Tian 2005].

imputed iteratively [Van Buuren 2012, pp. 108–116]. The FCS method was selected because it can be applied to a general pattern of missing data, which is the case of our example (see Table 1). The imputations are based on the *predictive mean matching* in which the values are imputed from donors selected according to the regression model [Van Buuren 2012, pp. 68–74]. *Predictive mean matching* ensures that imputed values are from the range of the variable of interest. In our case it means that imputed values are $x_{ik}^{m} \geq 0$. We are assuming that *income* is MAR. Thus, the imputation model should include the variables with which we are correcting for missingness. The general idea is to include all the potential predictors in the imputation model. However, for large datasets it might be at the very least not that efficient. The solution is to use only selected predictors which are considered to be related with the variable of interest and with the nonresponse on that variable [Van Buuren 2012, pp. 127–128]. The list of variables chosen as predictors for the imputation model for *real disposable income* is given on page 97.

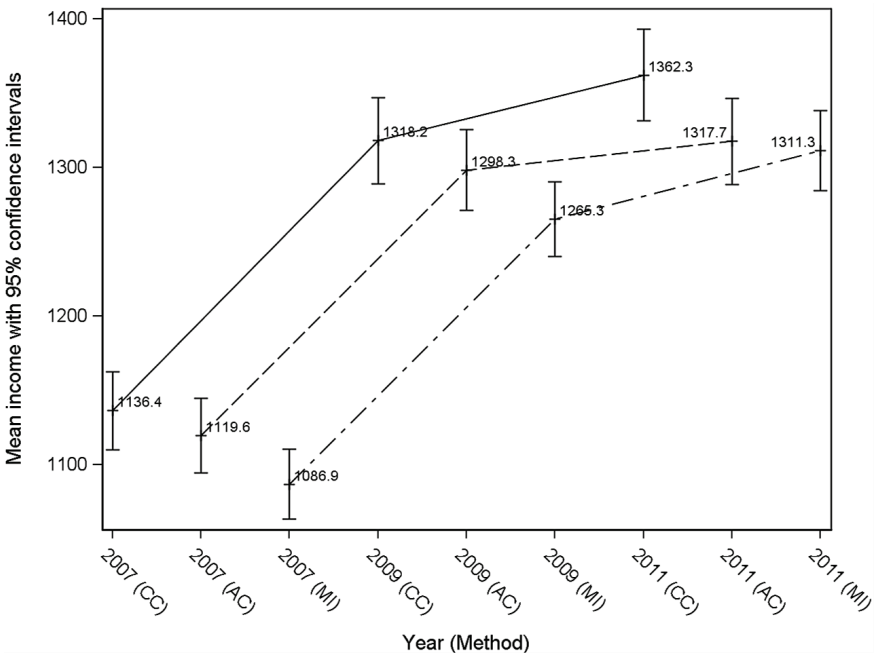

Figure 4. Point estimates of the mean income in 2007, 2009 and 2011 with 95% confidence intervals calculated for complete-cases (CC), available-cases (AC) and with use of multiple imputation (MI)

Source: own elaboration.

The results of the estimation of *real disposable income* in years 2007, 2009 and 2011 with use of *complete-case*, *available-case* and *multiple imputation* are given in Figure 4 presenting the point estimates with 95% confidence interval for the mean *real disposable income*. As we can see on the graph the traditional methods give higher estimates compared to the multiple imputation estimates. Assuming *income* is MAR on the given predictors, the traditional methods overestimate the mean *real disposable income* in each of the years studied. In this case we can conclude that the respondents with missing incomes are those with low incomes.

The method used in the above example considers two possible scenarios, namely that *income* is missing completely at random (*complete-case* and *available-case* analysis) and that *income* is missing at random (*multiple imputation*). If one considers that the missingness on *income* depends on its value (which is equivalent to is not missing at random assumption) further investigation would be required (see for example [Van Buuren 2012, pp. 88–93; Little, Rubin 2002, pp. 321––327, Zmyślona 2006]).

4. Conclusion

In conclusion let us assess the advantages and disadvantages of both traditional and modern approaches to the problem of missing-data. The main advantage of the traditional methods is in their simplicity which makes the analysis and interpretation straightforward [Little, Rubin 2002, p. 41]. However, in general with the use of traditional methods a researcher is unable to get at once unbiased, precise and efficient estimates (see [Laurens 1999, p. 39]). The reason is in the very restrictive assumptions that have to be made when using the traditional methods, like for instance the MCAR assumption. Because of these deficiencies the traditional methods cannot be considered as a remedy for the missing-data problem. However, they can be viewed as a good starting point for the missing data analysis or a part of the sensitivity analysis.

The modern methods for handling survey nonresponse are designed to give statistically valid and efficient estimates under the MAR assumption which is less restrictive that the MCAR assumption made in the case of a *complete-case* analysis [Molenberghs, Kenward 2007, p. 78]. Moreover, they offer the possibility to incorporate the knowledge about the missingness which is in the data set and which is possessed by the data collector [Rubin 1987, p. 15]. What is more, the modern

approaches allow us to model the missingness mechanism and check for the impact of posited, still empirically unverifiable assumptions on the results [Van Buuren 2012, p. 93]. From the theoretical point of view the main drawback of the modern methods lays in their sensitivity to the choice of imputation model. Still, one can use several imputation models and check the sensitivity of the substantive analysis to the choice of imputation model. While this makes the analysis more complex [Rubin 1987, pp. 17–18], which is an obvious disadvantage, the increase in complexity is significantly reduced by the fact that the current methods are widely available in statistical software [Molenberghs, Kenward 2007, p. 45].

References

Allan F.E., Wishart J., *A method of estimating the yield of a missing plot in field experiment work*, "Journal of Agricultural Science" 1930, Vol. 20, No. 3, pp. 399–406.

Allison P., *Missing data*, [in:] R.E. Millsap, A. Maydeu-Olivares (eds.), *The SAGE Handbook of Quantitative Methods in Psychology*, SAGE Publications, London 2009, pp. 72–89.

Allison P., *Multiple imputation for missing data: A cautionary tale*, "Sociological Methods Research" 2000, Vol. 28, No. 3, pp. 301–309.

Balicki A., *Metody imputacji braków danych w badaniach statystycznych*, "Wiadomości Statystyczne" 2004, No. 9, pp. 1–19.

Bracha C., *Metoda reprezentacyjna w badaniu opinii publicznej i marketingu*, Efekt, Warszawa 1998.

Carpenter J.R., Kenward M.G., *Multiple Imputation and its Application*, John Wiley & Sons, Chichester 2013.

Carpenter J.R., Kenward M.G., Vansteelandt S., *A comparison of multiple imputation and doubly robust estimation for analyses with missing data*, "Journal of the Royal Statistical Society: Series A" 2006, Vol. 169, No. 3, pp. 571–584.

Fisher R.A., *The Design of Experiments*, Hafner Press, New York 1971.

Graham J.W., *Missing Data. Analysis and Design*, Springer, New York 2012.

GUS (Polish Central Statistical Office), *Consumer Price Index*, http://www.stat.gov.pl/gus/5840_1638_PLK_HTML.htm (21.02.2013).

Heitjan F., Little R.J., *Multiple imputation for the fatal accident reporting system*, "Applied Statistics" 1991, Vol. 40, No. 1, pp. 13–29.

Kalecki M., *On the Gibrat Distribution*, „Econometrica" 1945, no 13(2), pp. 161-170.

Laurens J.P., *Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets*, Partners Ispkamp, Enschede 1999.

Little J.A., Rubin D., *Statistical Analysis with Missing Data*, John Wiley & Sons, Hoboken 2002.

Molenberghs G., Kenward M.G., *Missing Data in Clinical Studies*, John Wiley & Sons, Chichester 2007.

Paradysz J., Szymkowiak M., *Imputacja i kalibracja jako remedium na braki odpowiedzi w badaniu budżetów gospodarstw domowych*, "Taksonomia" 2007, No. 14, pp. 74–81.

Rubin D.B., *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, Hoboken 1987.

Schafer J.L., *Analysis of Multivariate Incomplete Data*, Chapman & Hall, London 1997.

Social Diagnosis 2000–2013, http://www.diagnoza.com/index-en.html (2.01.2012).

Szymkowiak M., *Badanie możliwości wykorzystania informacji pochodzących z rejestrów administracyjnych do kalibracji w krótkookresowej i rocznej statystyce przedsiębiorstw*, Zeszyty Naukowe Uniwersytetu Ekonomicznego w Poznaniu 227, Poznań 2012, pp. 140–156.

Tian L., *Inferences on the mean of zero-inflated lognormal data: The generalized variable approach*, "Statistics in Medicine" 2005, Vol. 24, pp. 3223–3232.

Van Buuren S., *Flexible Imputation of Missing Data*, Taylor & Francis Group, Boca Raton 2012.

White I., *Handling missing outcome data in randomised trials. Lecture 3: Multiple imputation*, unpublished course materials (14–15.03.2013), MRC Biostatistics Unit, Cambridge 2013.

Zdobylak J., Zmyślona B., *Analiza niepełnych danych w badaniach ankietowych*, [in:] W. Ostasiewicz (ed.), *Ocena i analiza jakości życia*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław 2004, pp. 269–323.

Zmyślona B., *Uwagi na temat własności estymatorów wyznaczanych na bazie niepełnych danych*, „*Ekonometria*" 2011, no 30, pp. 83–93.

Zmyślona B., *Zastosowanie modeli hierarchicznych w bayesowskim wnioskowaniu statystycznym w przypadku danych niepełnych*, "Ekonometria" 2006, nr 17, pp. 30–41.

PRZEGLĄD METOD ANALIZY NIEKOMPLETNYCH ZBIORÓW DANYCH WRAZ Z PRZYKŁADAMI ZASTOSOWAŃ

Streszczenie: Celem opracowania jest przegląd tradycyjnych (*usuwanie wierszy*, *usuwanie wierszy parami*, *imputacja pojedyncza*) i współczesnych (*metoda największej wiarygodności*, *imputacja wielokrotna*, *metody wagowe*) metod stosowanych wobec niekompletnych zbiorów danych oraz ocena ich użyteczności z punktu widzenia analiz statystycznych. W opracowaniu podano podstawowe pojęcia, a także opis podstawowych metod i algorytmów. Metody oceniono, uwzględniając własności estymatorów otrzymywanych na ich podstawie. Wybrana metoda (*imputacja wielokrotna*) zilustrowana została przykładem. Przegląd wskazuje na to, że metody współczesne mają przewagę nad metodami tradycyjnymi pod względem redukcji obciążenia, precyzji, a także efektywności estymatorów.

Słowa kluczowe: wzorzec brakujących danych, mechanizm brakujących danych, usuwanie wierszy, usuwanie wierszy parami, pojedyncza imputacja, metoda największej wiarygodności, imputacja wielokrotna, metody wagowe.