UDC 316.42 – 047.44

6. MODERN METHODS OF ANALYSING SOCIOLOGICAL DATASETS WITH MISSING VALUES

Uchkov A.K., Master's degree student, group 325901

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Subbotkina I.G. – Associate Professor

Annotation. This article discusses methods for handling missing data in analysis of surveys. It covers various imputation methods, classification of missing data, one of the approaches to evaluate imputation effectiveness and guidelines for method selection. Key considerations such as dataset structure, relationships among variables, and post-imputation analysis tools are highlighted.

Keywords. Missing data, data analysis, imputation methods, experimental effectiveness evaluation.

Data analysis obtained from surveys and measurements of socio-economic indicators describing the studied object or phenomenon involves processing large datasets based on a set of variables. However, often such variables contain missing values. There are many reasons why a dataset can have missing data – some of them are associated with the respondent himself, such as skipping large questions or questions that can be considered personal. Another group of reasons is connected to the technical part of research: data loss in the processes of sending responses to the analysis team, wrong data encoding, or surveys being cut short because of power outages.

Before figuring out how to deal with missing data, it's important to understand what type of missing data is present. Classifying missing data helps researchers identify patterns and biases in the dataset. Recognizing the type of missing data helps to interpret given data correctly and enhances the transparency and integrity of the research outcome [1].

In 1992, the following classification of missing data was proposed [2]:

1. MCAR (Missing Completely At Random) - a mechanism where the probability of missing data is the same for each record in the dataset. It appears when some unexpected event interrupts the survey, for example, losing the Internet connection. Also, sometimes respondents may skip some questions randomly, so there is no correlation between these empty values and other answers of that person.

2. MAR (Missing At Random) - a mechanism where data is missing not randomly, but due to certain factors. A missing value is classified as MAR if its probability can be calculated based on other available information in the dataset. For example, if information about gender, income and age is collected in a survey, a correlation may exist: if a person's gender is female, it's more likely to have missing data about age, and if a person's gender is male – more missing values of income may be present.

3. MNAR (Missing Not At Random) - a mechanism where the absence of values is associated with unknown factors. As a result, the probability of missing data cannot be expressed based on the information contained in the dataset. The presence of MNAR-type missing values in the dataset is a sign to the researcher that it is necessary to improve the quality of the survey.

To address missingness in data, researchers use the process of imputation. However, it should be noted that methods for imputing missing values are applicable only when data gaps are MCAR or MAR; using imputation to cover MNAR values could lead to the construction of a low-quality model because only MCAR and MAR values can be restored within the given dataset.

Today, there are numerous algorithms developed for solving the problem of imputation. It is important to understand the process of choosing the correct algorithm because this process would be based on the types of missing values in datasets and the dataset's structure. However, in practice, there are three basic approaches that are used when dealing with missing values [3]:

1. Deleting incomplete observations from the database. Observations containing missing values will be excluded from the analysis, even if missing variable isn't included in current analysis. This method is very easy to use, however, it affects data negatively, as removing observations leads to a reduction in the sample size, loss of information, and bias in the analysis results. This method can be applied without significant consequences if the missing values correspond to the MCAR category and the percentage of incomplete observations is relatively small (less than 5 %). Another approach is only to delete observations containing missing values in variables that are under consideration at the moment; if a record has a missing value in another variable, it won't be deleted. However, if applied multiple times, relationship between different results may be incomparable, as they will use different subsamples from the original dataset.

2. Weighting observations. This method allows preserving the required sample size while removing incomplete observations. A weight coefficient is assigned to each complete observation depending on the variable for which the structure needs to be preserved. Assigning a weight that is greater than one increases

the size of data sample, because observation will be accounted multiple times. The drawback of this method is that assigning weights may lead to a significant bias in parameter estimation.

3. Imputation of missing values. This is a group of different methods that replace missing values based on some rule or a set of rules. These methods can be divided into simple and complex ones. Complex methods can use either similar observations to fill in missing values (local methods) or they can rely on the entire dataset to restore missing values (global methods).

A method is considered simple if there is only one iteration in it. One of the first methods that have been used for a long time is a mean imputation: missing values are substituted by the mean value of variable. This, of course, can be done only if a variable is an interval variable; if it is an ordinal variable, the substitution by a median is used, and if it is a nominal variable – missing values are substituted with a mode instead. This method can only be applied to normally distributed data; otherwise, some of the values may be far from the mean, leading to distortion of the data structure. This method typically provides results with high prediction errors.

The HotDeck method restores missing values of a particular object's feature. For example, there are n observations and m variables in dataset. The method is based on the assumption that if objects are similar in terms of the values of the m-1 variables, they are also similar in terms of the variable m. Filling in the missing value of a variable for an incomplete object involves substituting the value of the same variable from the nearest complete object. Similarity measures can be computed for variables of all data types. The HotDeck method also can be combined with cluster analysis, when complete objects are clustered together. The disadvantage of this method is a large consumption of computing power to find a similar object or create a cluster [4].

In contrast to HotDeck imputation, which uses observations from the current data set, ColdDeck imputation uses external sources, usually results of a previous survey. It imputes missing values using reported values from donors. Because the results of previous research are needed, this method is mostly used in *longitudinal studies.*

The regression analysis method constructs a multiple linear regression model where the dependent variable is the column containing missing values, and the independent variables are columns with complete values. To estimate the coefficients of the regression equation, least squares method is used. The unknown missing value is calculated by substituting values of the current record into the regression equation. The drawback of this method is the need of additional data preprocessing: only variables highly correlated with the dependent variable should be included in the set of independent variables. Also, variables highly correlated with the dependent variable frequently contain missing values too. This method is useful if a linear correlation is present, but may give imprecise results with non-linear correlations.

These are most used simple methods of imputation. In comparison to them, complex methods use multiple iterations to get more precise results. It should be mentioned, however, that both simple and complex methods can be optimal, it depends on the dataset and target variables.

The ZET algorithm belongs to the class of complex local methods. It is based on the following assumptions:

1. Redundancy. The assumption is that there are observations in the population being studied that are similar to each other; there are also some variables that similar to each other.

2. Analogy hypothesis. The assumption is that if objects are similar in all variables, except one, then they are similar in that one too. It is also used in HotDeck method.

3. Local competence. The assumption is that to predict missing values, not the entire matrix is used, but only the part that consists of elements of close rows and similar columns. The «competent» part should not contain missing values.

The ZET algorithm includes the following steps: selection of the «competent» part to fill in the missing value; calculation of coefficients for the equation used to forecast the missing value; calculation of the forecasted value [5].

The Bartlett algorithm is a global complex method that is based on the regression method and includes two iterations:

1. Substituting initial values for missing entries.

2. Conducting covariance analysis of the target variable and a dichotomous indicator of observation completeness.

The indicator's value is always equal to 0, unless the target variable is empty in the record; then, the value equals to 1. The drawback of the Bartlett method is similar to that of the regression analysis method: it is associated with the assumption of linear dependence between variables, which is not always observed in practice.

The expectation-maximization is another global complex method, where an algorithm builds a missing data generation model with conclusions based on the maximum likelihood function. Each iteration of the algorithm consists of two steps.

1. The E-step (expectation): the expected value of the likelihood function is calculated, with hidden variables treated as observable.

2. The M-step (maximization), the maximum likelihood estimation is computed, thereby increasing the expected likelihood calculated in the E-step. This value is then used for the E-step in the next iteration.

The algorithm continues until the convergence. The drawback of the method lies in difficulty of constructing the missing data generation model [3].

The resampling algorithm is an iterative method that can be implemented in two modifications. In the first modification, missing values of incomplete observations are randomly replaced with corresponding values from complete observations in the original dataset, and then a regression equation is constructed. In the second variant, the regression equation is obtained from the complete submatrix. Random variable values in the first case and the low power of the set of complete variables in the second case can lead to incorrect results.

Multiple imputation of data is the most common method for filling in missing values in a sociological practice today. The essence of the method is to substitute multiple values for each missing value, i.e., k datasets or k matrices are formed. Then, the missing value is replaced with the mean value calculated from all constructed models. Each set is obtained using one of the following models: predictive, propensity degree, or discriminant. The drawbacks of the method include significant time and computational costs compared to any of the methods discussed above [3].

Neural networks are used to solve a wide range of tasks, including clustering, pattern recognition, optimization, and others. Neural networks can also be applied to forecast missing values.

In the realm of data analysis and research, choosing the correct method for handling missing values can be challenging, as there is a lot of available options. The complexity lies in selecting the most suitable approach for each unique dataset and research objective. However, there are some guidelines that can be applied in most situations.

If there is less than 5 % of missing data, multiple imputation methods may not offer significant advantages, so it can be useful to use single imputation approach or to use deletion methods. But if percent of missing data lies in range from 10 % to 40 %, a bias is more likely, so the use of multiple imputation techniques is preferred. When over 40 % of data is missing, any kind of methods becomes unstable, and can be used only to generate hypothesis. If that much of data is missing, researcher needs to use his own knowledge of data field to investigate the data and find variables that may be analyzed. If missing percent data is less than 10 %, but more than 5 %, all variables may be analyzed, but the choice still should be based on theoretical knowledge about the subject of research.

Another important thing to find out if the data is MCAR, MAR or MNAR. Little's test of missingness [6] can be performed to decide if the given missing data can be classified as MCAR. However, if the missing data cannot be classified as MCAR, researcher should find out if the missing data is connected to another variable from the dataset. If it is true, then the missing values may be classified as MAR. In both of that cases, different methods can be used further on. In the third case, if these values are MNAR, the dataset should be investigated, like in the case if more than 40 % of data is missing [1].

Nowadays, a lot of data processing software packages contain numerous methods, and therefore another idea to find out optimal algorithm for handling missing data in each scenario can be formalized. At the start, the structure of the sample, the relationships between its variables, and the research tools that will be applied after missing value imputation should be analyzed. Then, when possible algorithms list is shortened, the effectiveness of a particular method can be established experimentally:

1) A dataset of complete records is formed by excluding incomplete observations from consideration.

2) Certain elements from the table are removed, in correlation to missing values from the original table.

3) The missing values are predicted using different methods.

4) The total relative errors are calculated for each method.

The idea is simple: if we know each value of each variable from the subset, then different methods can be compared by creating artificial empty values and replacing them with predicted ones. The method with the minimum total relative error is considered the most effective. Since there are many imputation methods available, it is practical to test the methods implemented in the data processing package used by the researcher [3]. To develop the most suitable simulation, two main factors should be considered: what types of records are most likely to have missing values, and what variables in these records are most likely to be missing. For example, if the values are deleted at random from the randomly selected records, then this missingness can be classified as MCAR, and the simulation result may not provide the best solution for MAR and MNAR values.

References:

1. Mirzaei, A. Missing data in surveys: Key concepts, approaches, and applications / Mirzaei, A. [et al.]. – Research in Social and Administrative Pharmacy – 2022. Vol. 18, iss. 2,

2. Little, R. J. A. Statistical Analysis with Missing Data. / R. J. A. Little, D. B. Rubin. – John Wiley & Sons – 2002.

3. Fomina, E. E. Review of software and methods for recovering missing values in sociological data sets / E. E. Fomina. – Humanities bulletin of BMSTU. – 2019. Vol. 4, iss. 78.

4. Jehanzeb, R. C. A Review of Missing Data Handling Methods in Education Research. / R. C. Jehanzeb. – Review of Educational Research. – 2014. Vol. 84, No. 4, pp. 487 – 508.

5. Algorithm ZET [Electronic resource]. – Mode of access: https://miest.narod.ru/iissvit/rass/vip39.pdf. — Date of access: 20.03.2024.

6. Little, R. J. A. 1988. A test of missing completely at random for multivariate data with missing values. / R. J. A. Little. – Journal of the American Statistical Association. – 1988. Vol. 83, iss. 404.