Categorical data imputation under MAR missing scheme

Pavel Zimmermann[1], Petr Mazouch[2], Klára Hulíková Tesárková[3]

Abstract. Traditional missing data techniques of imputation of the MAR (missing at random) schemes focus on prediction of the missing value based on other observed values. In the case of continuous missing data the imputation of missing values often focuses on regression models. In the case of categorical data, usual techniques are then focused on classification techniques which sets the missing value to the 'most likely' category. This however leads to overrepresentation of the categories which are in general observed more often and hence can lead to biased results in many tasks especially in the case of presence of dominant categories.

We present original methodology of imputation of missing values which results in the most likely structure (distribution) of the missing data conditional on the observed values. The methodology is based on the assumption that the categorical variable containing the missing values has multinomial distribution. Values of the parameters of this distribution are than estimated using the multinomial logistic regression.

Keywords: Missing data, Categorical data, Multinomial regression.

JEL Classification: C35
AMS Classification: 62J99

1 Introduction

Popular methods for a completion of (individual) observation as for example mean imputation, regression imputation or maximal likelihood imputation are usually focused on imputation of a continuous variable. Those methods mostly classify the missing values as "most likely" or "expected" values. Overview of those methods can be found for example in [5]. List of methods for imputation of categorical variable is less extensive. In the case of categorical data, usual techniques are then focused on classification techniques which sets the missing value to the 'most likely' category (see [6]). This however leads to overrepresentation of the categories which are in general observed more often and hence can lead to biased results in many tasks especially in the case of presence of dominant categories.

The aim of the paper is to introduce multinomial logistic regression as very effective tool for missing data imputation. Motives for using this technique could be described by the following three requirements:

1. to impute data set in form which can be re-used for variety of different analysis and applications; this means single imputation is required,
2. to impute data in the most detailed level; optimally on individual observation level
3. to impute data in a way that will respect "expected" ratios of categories in general.

In the following text the methodology and its specific features will be described.

2 Missing data typology

In this article the widely renowned typology of missing data structures developed in [4] will be adopted. Rubin considered the missingness as a probabilistic phenomenon, i.e. a set of random indicator variables R indicating non-missingness of a particular observation was considered. Also the partition of the complete dataset Y_{com} into set of observed values Y_{obs} and set of missing values Y_{mis}, i.e.

$$Y_{com}=(Y_{obs}, Y_{mis})$$

was considerd. Missing data are called missing at random (MAR) in the case where the distribution of the missingness does not depend on Y_{mis}, i.e. when

$$P(R|\ Y_{mis}) = P(R|Y_{obs}).$$

[1] University of Economics, Prague, Faculty of Informatics and Statistics, Department of Statistics and Probability, nam. W. Churchilla 4, Prague 3, zimmerp@vse.cz

[2] University of Economics, Prague, Faculty of Informatics and Statistics, Department of Economics Statistics, nam. W. Churchilla 4, Prague 3, mazouch@vse.cz

[3] Charles University in Prague, Faculty of Science, Department of Demography and Geodemography, Albertov 6, Prague 2, klara.tesarkova@gmail.com

This is the case where ‘MAR allows the probabilities of missingness to depend on observed data but not on missing data’. A special case of the MAR is then MCAR (missing completely at random), where the probabilities of missingness do not depend on the observed data either:

$$P(R|Y_{com}) = P(R).$$

If MAR is violated, data are missing not at random (MNAR).

The task solved within the paper

In this article a methodology for a specific task is developed which can be however reused in many similar tasks. Conditions of the solved task could be described as:

1. We have a univariate pattern of categorical data, i.e. data where several variables are completely observed (X_{obs}) and one variable contains missing values. This can be schematically expressed as in [5]:

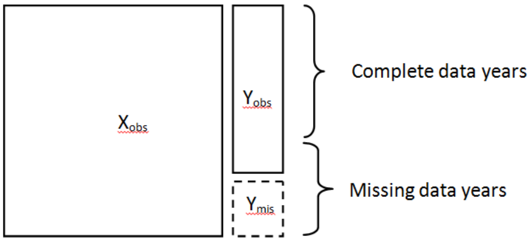

Figure 1 Data set structure

2. Data are observed over certain period of time and for some (major) part of the period data are either complete or with negligible amount of missing data. These years will be referred as ‘complete data years’.
3. For the outstanding years, the amount of missing data is rather large and MAR is not guaranteed. These years will be referred as ‘missing data years’.
4. The trends observed during the complete data years are relevant for the predictions for the missing data years.
5. Observations are assumed independent.

The time structure of data set according to missing data

From the time point of view three types of missing data position could be distinguished. The first is situation where we have complete information from some moment (year) but before this time missing data occur. This situation illustrates Figure 3a. In such a situation the aim is to reconstruct data before some point in time.

The second example is situation where data are complete, however, from some moment in time some (or all) data are missing. This situation is in Figure 3b and the aim in such a situation is to estimate the missing information for that period after any concrete moment. The third type is a situation where we have missing data “in the middle” of the time period, i.e. for some (limited) period of time the information is partially or completely missing. Figure 3c describes that situation. The aim is to bridge this part, estimate the missing data respecting trends before and after this missing period.

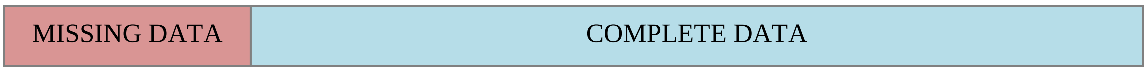

Figure 3a Structure of the data set – missing data at the beginning of the analyzed time period

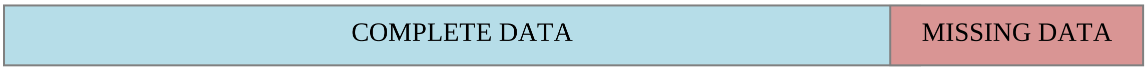

Figure 3b Structure of the data set – missing data at the end of the analyzed time period

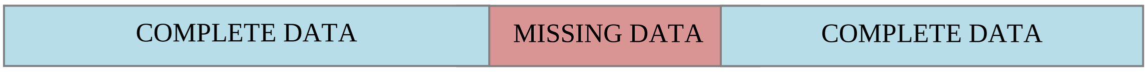

Figure 3c Structure of the data set – missing data for a limited time during the analyzed time period

3 The imputation algorithm

In the following text, we will mean by **determinants** the original or rediscretized variables that have a significant impact on the distribution of the variable containing missing data. (The significance is measured over the years with complete data.) By **profile** we then mean a group of data with the same combination of values of the determinants. We will assume time points t for which complete data years are the time points $t \le c$ and missing data years are $t > c$.

The basic steps of our imputation algorithm:

1. Find determinants of the missing data structure within the observed variables.
2. Define profiles of observations with missing data based on the values of the observed determinants. These profiles have different distribution of the incomplete variable.
3. Estimate probabilities of each category of the missing variable for each profile.
4. Based on the probabilities, find "appropriate" count of missing observations of each category in each profile and distribute these counts to each individual in the profile.

3.1 Multinomial logistic regression application

If we assumed that data are independent (independence of the observations) the categorical variable containing the missing values (Y) follows for a given profile the multinomial distribution. This fact immediately suggests using the multinomial logistic regression on the complete data years ($t \le c$) as the methodology for finding the determinants (X or subset of X) of the structure of Y (as the response variable) and predicting the expected probabilities of each category of the response variable for each profile of data at each time point (for both $t \le c$ and $t > c$). This requires assessing the time variable as covariate and assuming some (possibly polynomial) trend. That is the probability distribution of the categories of Y for each profile in each year $P(Y|X,t)$ is fitted as the outcome of the regression analysis (steps 1-3 of the above outlined imputation algorithm).

3.2 Partially missing data

Based on the above described analysis we obtain the predicted distribution of the variable containing the missing data (Y) also for the years containing missing data ($t > c$) for each profile and each time point (conditioning on X and t will be leaved out in this section for simplicity). However, for these years we may have some amount of observed data (supposing partially missing data in the data set). Therefore we can estimate two distributions of missing values, first based on complete data years and second based on missing data years:

1. First distribution as the prediction based on the complete data years $P(Y = i)$ for each category $i = 1,\dots,k$ and a given profile and each time point.
2. Second distribution fitted based on the observed data $P(Y = i|\ R=0)$ in the missing data years, i.e. distribution conditional on the fact that an observation is not missing.

Besides these distributions, we can also estimate the probability of missing values $(P(R=0))$. The (marginal) distribution $P(Y=i)$ equals

$$P(Y=i) = P(Y=i,R=0) + P(Y=i,R=1),$$

where $P(Y=i,R=0)$ (or $P(Y=i,R=1)$) is the (joint) probability that the observation is certain category and is missing (or is not missing respectively) which equals

$$P(Y=i,R=0) = P(Y=i|R=0)\ P(R=0) \text{ and}$$

$$P(Y=i,R=1) = P(Y=i|R=1)\ P(R=1).$$

We can write for the distribution of the observations that are missing (i.e. for which we already know that R=0) as:

$$P(Y=i|R=0) = [\ P(Y=i) - P(Y=i|R=1)\ P(R=1)\ ]\ /\ P(R=0).$$

3.3 Finding the appropriate count of missing observations of each category

Let us assume one particular profile of the data in a given year. Based on the above described regression analysis we can get the estimated distribution of the categories of Y for the missing observations, denoted as $P(Y=i|R=0) = p_i,\ i=1,\dots,k$ for this given profile and year. Furthermore we know that in this profile and year, there is certain amount of missing data n. Given the probability distribution of the categories of the response variable and the number of missing observations we still need to determine how many of the missing observations correspond with each category (step 4 of the above outlined imputation algorithm). Note that it is required that the missing values are imputed on the individual level and hence we need to determine counts (integers) of missing observations for each educational category.

Normally the expected value would be the first choice for the predictions as it yields predictions with the lowest least square error. The expected value of the multinomial distribution in a particular category *i* is simply the count of missing observations (in the particular profile in the particular year) times its probability, i.e.

$$E(y_i) = n\,p_i\,, \quad i=1,\ldots,k.$$

However, the expected values are generally real numbers (not necessarily integers). Therefore we suggest using the maximum likelihood criterion where the maximization is performed only on the discrete (integer) numbers. This means finding such y_i , $i=1,\ldots,k$ that the joint distribution $P(y_1,y_2,\ldots,y_k\ |p_1,p_2,\ldots,p_k,\ n)$ is maximized. This in fact means we are looking for the mode of the multinomial distribution.

Mode of the multinomial distribution

There is no closed form formula for the mode of the multinomial distribution. There are however several iterative algorithms developed for this task. See for example [2], [1] or [3]. In our computations we selected the **Finucan's algorithm** published in [1].

Distribution of estimated data on the individual level

Having found the mode of the multinomial distribution for a particular profile we have a vector of counts (integers) of missing values of each category of the variable of the concern which has the highest probability. Within the profile, these counts may be 'assigned' randomly to the individuals as all individuals of the given profile have the same probability vector $p_i,\ i=1,\ldots,k$ of being in *i*-th category.

4 Discussion

The proposed method of estimation of missing data could be used in many spheres of application. In this paper we demonstrated the algorithm on (completely or partially unknown) education structure of a population. Education attainment could be taken as a typical example of categorical data. Moreover, when studying the population, this type of data is relatively often incomplete. Other example could be e.g. the marital status, age profile, etc.

The described algorithm is based on the assumption of continuous trend in the data within the missing data years. It corresponds with situation where data are missing because of some administrative changes etc. which does not affect the trend in the data. Application of the described method in situations where this condition is not fulfilled (e.g. where the missingness of the data is at least partially related to some changes affecting also the long-term trend – wars, etc.) would mean some sort of simulation of "unaffected" development – how the structure (partially or completely missing) would have developed if there had not been any interruption of the trend.

Furthermore the estimates of the differences between the distributions $P(Y=i)$ and $P(Y=i|R=1)$ may suggest the (non)randomness in missingness 'mechanism.' The results in integer form have the advantage that it allows imputing data on individual level. (Such imputation is however not unique).

5 Illustrative example

As an illustrative example we assumed educational attainment of a studied population as the variable containing missing data. For simplicity we assume that data are complete up to 2009 and data are completely missing in 2010 and 2011. The variable has 3 categories (low, middle and high education). There is only one other variable (*X*) which is the gender. Therefore we only have two profiles in each year. The probabilities estimated with the multinomial logistic regression are displayed in the Figure 4.

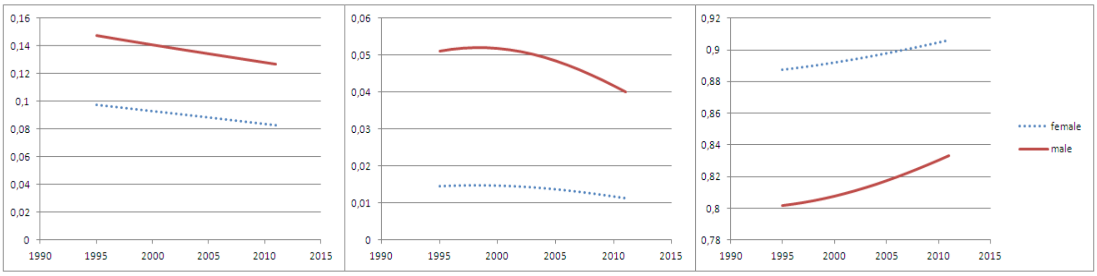

Figure 4 Estimated probabilities of the educational categories.

For illustration we will assume only the year 2011 where we know that we assume that we are missing 100 females and 50 males. The probabilities, expected counts and the mode of the multinomial distribution are contained in the Table 1.

2011	Count missing	Probabilities			Expected counts			**Mode**		
		Low	Middle	High	Low	Middle	High	**Low**	**Middle**	**High**
Female	100	0,906	0,083	0,011	90,62	8,26	1,12	**91**	**8**	**1**
Male	50	0,833	0,127	0,040	41,65	6,35	2,00	**42**	**6**	**2**

Table 1 Estimated probabilities, expected counts and the mode of the multinomial distribution.

Based on these results, we will impute 91 females with low education, 8 females with the middle education and 1 female with high education and analogously for males. As there is no more information available, these counts are distributed within each profile randomly.

6 Conclusion

Aim of this paper was to introduce multinomial logistic regression as very effective tool to missing data imputation. To the authors' knowledge the combination of the multinomial regression and mode searching algorithm was used for the first time for the missing data imputation task. The outcome of the proposed algorithm follows expected structure of the variable containing the missing values.

As a by-product the outcomes of the intermediate steps of the algorithm may be used for further analyses such as analyses of the dependencies (determinants) of the variable of our concern, or analysis of the missingnes mechanism.

Future steps in the research will be to proof this method in some other practical situation. Demographic data (with incomplete information about the education attainment occurring in the latest years of the involved time period –as in the Figure 3b) were used for the very first verification of the model and first results seem to be acceptable. Next part of the research will be to find more datasets with missing data, both MAR and MNAR and with different structure of missing data from the time point of view (length of missing, time of missing) and to prepare more detailed analysis of complemented data files.

Acknowledgements

The article was written with the support provided by the Grant Agency of the Czech Republic to the project no. P404/12/0883 "*Generační úmrtnostní tabulky České republiky: data, biometrické funkce a trendy*".

References

[1] Finucan H. M., The Mode of a Multinomial Distribution, *Biometrika*, Vol. 51, No. 3/4, Dec., 1964 (pp. 513-517)

[2] Johnson, Norman Lloyd, Samuel Kotz, and Narayanaswamy Balakrishnan. *Discrete multivariate distributions*. Vol. 165. New York: Wiley, 1997.

[3] Le Gall F., Determination of the modes of a Multinomial distribution, *Statistics & Probability Letters*, Volume 62, Issue 4, 1 May 2003, Pages 325–333, http://dx.doi.org/10.1016/S0167-7152(02)00430-3

[4] RUBIN, D. B. Inference and missing data, *Biometrika* (1976), 63 (3): 581-592.

[5] Schafer, Joseph L., and John W. Graham. "Missing data: Our view of the state of the art." *Psychological methods 7.2* (2002): 147-177.

[6] Sentas, Panagiotis, A. Lefteris, and Ioannis Stamelos. "Multiple logistic regression as imputation method applied on software effort prediction." *Proceedings of the 10th International Symposium on Software Metrics*. 2004.