Gazi University Journal of Science
GU J Sci
25(4):869-873 (2012)

ORIGINAL ARTICLE

Application of Multiple Imputation Method for Missing Data Estimation

Gazel SER[♠]

Yuzuncu Yil University, Faculty of Agriculture, Biometry-Genetic Unit, Van, TURKEY

Received:17.11.2011 Revised:05.03.2012 Accepted:17.05.2012

ABSTRACT

The existence of missing observation in the data collected particularly in different fields of study cause researchers to make incorrect decisions at analysis stage and in generalizations of the results. Problems and solutions which are possible to be encountered at the estimation stage of missing observations were emphasized in this study. In estimating the missing observations, missing observations were assumed to be missing at random and Markov Chain Monte Carlo technique and multiple imputation method were applied. Consequently, results of the multiple imputation performed after data set was logarithmically transformed produced the closest result to the original data.

Key Words: *Multiple imputation, Missing data, Milk yield*

1. INTRODUCTION

Missing observations were quite often encountered in scientific researches. In case of missing data, tree major problems emerge. The first one occurs as data loss, second problem arises from the calculation and analysis stage due to the disorder of the structure, the third and most important problem causes bias results arising from the systematic difference between the observed and unobserved data [1]. One of the methods utilized for the estimation of missing observation and solution of the above mentioned problems is the Multiple imputation method. Varied solution and imputation methods were developed for the missing data problem. It is needed to identify the missing data mechanism before such solution and imputation methods. Because the decision which solution and imputation method would be applied to the data set depends upon the missing data mechanisms [2]. Little and Rubin classified those mechanisms under three basic categories. Accordingly; missing observation mechanism is missing completely at random (MCAR) if missing observations is independent of both observed and unobserved values, missing is at random (MAR) if it is independent of unobserved values and dependent of observed values and it is missing not at random (MNAR) if it depends on both observed and unobserved values [3-5].

It is necessary to achieve specific properties to be able to apply the Multiple Imputation method. Accordingly; missing data should have the MAR structure, model which is used to obtain the values imputed should be formed "correctly" and the model used for the analysis should comply with the model used at the stage of imputation [6].The multiple imputation method is a procedure of eliminating losses in the data set with two or several acceptable values representing probability distribution. It is necessary to determine/estimate the probability distribution regarding the complete data (observed or unobserved) in order to do a multiple imputation. The statistical inference via multiple imputation requires three basic steps. As for the first one, the missing observation in the data set were estimated *m* times and *m* complete data sets are formed, and this step is also known as the imputation step. At the second *m* complete data sets are analyzed through *m* standard

[♠]Corresponding author, e-mail: gazelser@yyu.edu.tr analyses methods, this step is called as analyzing step, at the third step the results obtained from m analysis were incorporated and inferences are made [7].

In this research, milk yield containing missing observation with repeated measure structure was included as the dependent factor in the model and estimation was made on the 205 missing observations contained in the data set. The missing observations were assumed to be MAR and missing observations were estimated through Multiple Imputation (MI) method using the MCMC technique in the estimation of missing observations.

2. MATERIALS AND METHODS

The animal material of the research included 41 Akkeçi goats between 1-5 years of age. Goats raised on the farm of Agriculture Faculty of Ankara University. Milk yield was controlled for at 14-day intervals during the lactation, measuring day was established by performing twice a day including morning and evening milking.

2.1. The Missing Observation Case and Missing Observation Estimation

The repeated measure values for i individual $(i = 1,2,...,N)$ are shown as $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in})$ and covariates values are shown as $X_i = (X_{i1}, X_{i2}, ..., X_{ip})$. Some repeated measures are separated as response variable (Y_i), $(Y_i^{(0)}, Y_i^{(m)})$ when they are not observed. While $(Y_i^{(0)})$ symbolizes the response vector of the observable i individual $(Y_i^{(m)})$ indicates the unobserved response vector of i individual. The missing case of observation vector $n \times 1$ dimensional indicator vector of each individual is indicated as $R_i = (R_{i1}, R_{i2}, ..., R_{in})$. If the response variable was observed as (Y_i) in other words there is no missing observations, it is considered as $R_i = 1$. However if (Y_i) contains missing observation then it is considered as $R_i = 0$ ([8],[9]). MCAR is called as the case that the possibility of missing observation which is intermittently available in the responses, found completely independent from all missing responses $(Y_i^{(m)})$, all observed responses $(Y_i^{(0)})$ and all the covarite (X_i). In other words, it is the case when it is independent from the (R_i), $(Y_i^{(0)})$ and $(Y_i^{(m)})$ in studies. MCAR mechanism is indicated as,

$$\Pr(R_i \mid Y_i^{(0)}, Y_i^{(m)}, X_i) = \Pr(R_i \mid X_i) \text{ [10].}$$

In case of MAR, the probability of the missing observation depends on the observed responses. When the case of missing observation is conditionally independent from $(Y_i^{(m)})$ if (R_i) s and $(Y_i^{(0)})$ s are given, the probability of MAR case is indicated as $\Pr(R_i \mid Y_i^{(0)}, Y_i^{(m)}, X_i) = \Pr(R_i \mid Y_i^{(0)}, X_i)$. Missing value distribution of each individual in MAR at $(Y_i^{(0)})$ is same with $(Y_i^{(m)})$ complete observation distribution [11]. For example, if the responses have multivariate normal distribution, the estimation of the missing values depends on the conditional average of $(Y_i^{(m)})$ when the $(Y_i^{(0)})$ is provided [8]. The case of MNAR is related to the conditional distribution of (R_i) when $(Y_i^{(0)})$ is given; and indicated as $\Pr(R_i \mid Y_i^{(0)}, Y_i^{(m)}, X_i)$ MNAR also refers to the missing not at random [12]. The assumption that missing observation is MAR is applicable in the MI technique. In other words missing observations (Y_{mis}) depend on the observed values (Y_{obs}). However, conditions on the observed values (Y_{obs}) are not performed for the missing observations (Y_{mis}). A model $\Pr(Y \mid \theta)$ is taken into consideration for Y data. θ takes place as parameter vector in the model. Missing observations need an independent prior distribution. $Y_{mis}^1, Y_{mis}^2, ..., Y_{mis}^m$ m sets of the imputation values are obtained through a Bayesian procedures if a prior distribution is given for θ and is removed independent from the *posterior* estimator,

$$\Pr(Y_{mis} \mid Y_{obs}) = \int \Pr(Y_{mis} \mid Y_{obs}, \theta) \Pr(\theta \mid Y_{obs}) d\theta \text{ (1).}$$

In practice, calculations made are a MCMC process, if parameter $Y_{mis}^{(t)}$ and $\theta^{(t)}$ values are given in MI t^{th} iteration. Random variables are removed from the probability distributions by the help of Markov chains. A Markov chain is a series of random variables, each individual value available in the distribution depends on the previous value in the series. In the MCMC simulation technique, a long-enough chain is formed and the stability of the distribution is ensured. The values by drawing random values from the conditional distributions as follows;

Step 1: $Y_{mis}^{(t+1)} \sim \Pr(Y_{mis} \mid Y_{obs}, \theta^{(t)})$

Step 2: $\theta^{(t+1)} \sim \Pr(\theta \mid Y_{obs}, Y_{mis}^{(t+1)})$

The Step (1) is the imputation step while the step (2) is the posterior or parameter step. Let the parameter estimate at the t step be $\theta^{(t)}$. At the imputation step, sample estimate is made at random for the missing data from $P(Y_{(mis)} \mid Y_{(obs)}, \theta^{(t)})$ distribution. Estimates obtained from the Step 1 that is shown as $Y_{(mis)}^{(t+1)}$. Those values are replaced at the posterior step and parameter estimate indicated as $\theta^{(t+1)}$ is made with the $P\left(\theta \mid Y_{(obs)}, Y_{(mis)}^{(t+1)}\right)$ probability distribution. Markov chain which consists of such estimates and converges to $P\left(Y_{(mis)}, \theta \mid Y_{(obs)}\right)$ distribution is formed $\left(\left(Y_{(mis)}^{(1)}, \theta^{(1)}\right), \left(Y_{(mis)}^{(2)}, \theta^{(2)}\right), \left(Y_{(mis)}^{(3)}, \theta^{(3)}\right) ....\right)$ ([4], [7], [13]).

3. RESULTS

The analyses were carried out using the multiple imputation module in the SPSS 17.0 software program. The Table 1 provides mean and standard deviation values obtained after being processed with both original and log-transformation of the milk yield.

Table 1. Descriptive statistics for properties of the milk yield.

	Missing				Std.
Variable	N	Percent	Valid N	Mean	Deviation
Milk yield	205	35.7%	369	606.233	340.698
Log of milk yield	205	35.7%	369	6.224	0.650

In the table, since the dependent variable has a repeated measure characteristic, total (41×14=574) measure values were obtained. The Linear Regression was used as the imputation model also time and age effects were included in the model.

The measure values obtained from the dependent values contain 205 missing observations. The multiple imputation results for the milk yield are provided in Table 2.

Table 2. Multiple imputation results for properties of the milk yield.

Data	Imputation	N	Mean	Std.Deviation	Minimum	Maximum
Original Data		369	606.233	340.698	100.000	1800.000
Imputed Values	1	205	576.834	365.742	-268.616	1618.575
	2	205	512.900	332.825	-317.353	1279.858
	3	205	559.466	353.979	-608.177	1561.640
	4	205	470.196	382.298	-434.966	1628.470
	5	205	551.524	351.691	-279.713	1332.737
Complete Data	1	574	595.733	349.814	-268.616	1800.000
After Imputation	2	574	572.899	340.570	-317.353	1800.000
	3	574	589.530	345.919	-608.177	1800.000
	4	574	557.648	361.713	-434.966	1800.000
	5	574	586.694	345.356	-279.713	1800.000

The table shows the imputed values of 205 observations in the original data and complete data obtained from the imputed and original data combination. In the research, 205 missing value estimation was carried out at 5 imputation step and total 1025 value imputation was performed. In the imputed values, lower value imputation was performed in comparison with the original data mean. The lower imputations were performed in comparison with the original data particularly at the 2nd and 4th steps however imputations which were closer to the original data were performed at 1st, 3rd and 5th imputation steps. Results which were partly closer to the original data were obtained in the complete data. However, imputations lower than the original data were performed for the maximum values while negative value imputation was performed for the minimum values. The maximum values for each imputation are considerably lower than for the original data. The distribution of milk yield tends to be right-skew, so this could be the source of the problem. The data was logarithmic transformed for the solution of this problem [14]. After logarithmic transformation was accordingly applied to the values of milk yield, missing observation imputation was performed by reapplying the MI method and the results obtained are provided in Table 3.

Table 3. The multiple imputation results logarithmic transformation applied that properties of the milk yield.

Data	Imputation	N	Mean	Std.Deviation	Minimum	Maximum
Original Data		369	6.225	0.650	4.605	7.495
Imputed Values	1	205	6.208	0.652	4.862	7.953
	2	205	6.183	0.630	4.703	8.078
	3	205	6.255	0.588	4.734	8.035
	4	205	6.224	0.638	4.741	7.748
	5	205	6.137	0.544	4.644	7.380
Complete Data	1	574	6.218	0.650	4.605	7.953
After Imputation	2	574	6.209	0.643	4.605	8.078
	3	574	6.235	0.628	4.605	8.035
	4	574	6.224	0.645	4.605	7.748
	5	574	6.193	0.615	4.605	7.495

In the Table 3, when the imputed values of 205 observations were assessed, it was found that the averages of the imputed values at other imputation steps expect for 5th Imputation step was closer to the original data average. The closest average estimation was obtained at 4th imputation step in the imputed values and complete data, and it was followed by respectively 1,3,2 and 5 steps. However, the negative value problem was eliminated at the lowest values and values equal to the original data were obtained. 3 pie charts summarizing the missing values in the data set were provided in Figure 1.

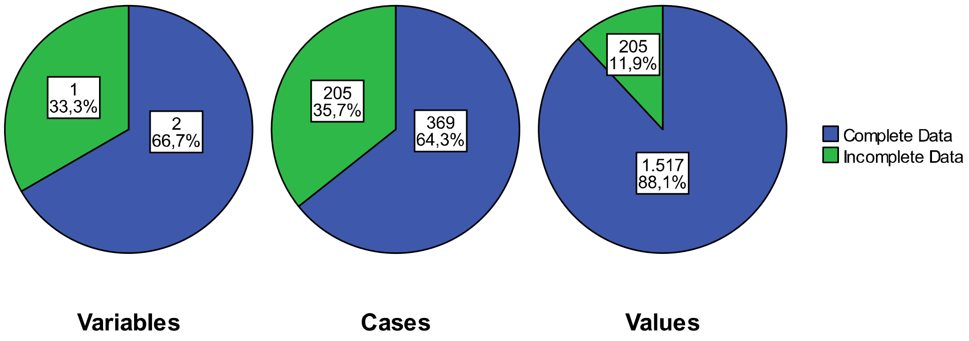

Figure 1. Overall summary of missing values.

Missing observation case of 3 variables in the data set was shown in *variable* chart. Accordingly, 2 of 3 variables, in other word 66,7% segment thereof, have not missing observation, only 1 variable contains missing observation. This variable constitutes the segment of 33.3%. In the *cases* chart, missing observation was found in the 35.7% segment of the observation number of the milk yield while 64.3% segment of the same does not contain missing observation. In the *values* chart, 205 observation values of 1722 (cases×variables) values are missing (574×3=1722-205=1517).

DISCUSSION

The multiple imputation method which may be applied to eliminate the missing data is a method based on the estimation of the missing data depending on the variable and/or observation value containing complete data in the sample. The statistical solutions is performed by the aid of new data set completing the missing data with estimated values [2]. The multiple imputation method provides an advantage when used under the MAR assumption. In this research, the problems arising from the use of multiple imputation method and solution points thereof were emphasized. The determination of the manner to find the missing observation in the dependent variable is quite important for establishing the imputation method. MI method used under the MAR condition produces better results. Also, the fact that negative values are found in the imputed values or the imputed values in the highest values is imputed quiet differently from the original date (e.g. performance of imputation much lower than the original data) indicates a problem about the distribution of the data. In such case, the appropriate transformation (e.g. logarithmic) process should applied to the main data and problem concerning the distribution should be eliminated [14]. The imputation step which has the closest value to the original data should be preferred when the data obtained from the respective imputation is chosen [15].

As a result, MI method correctly determined the structure of the missing observation. It is necessary to prefer the imputation that is the closest to the original data according to the results of imputation steps acquired using MI method.

REFERENCES

[1] Sartori, N., Salvan, A., Thomaseth, K., “Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose”, ***Computational Statistics & Data Analysis,*** 49, 937-953 (2005).

[2] Bal, C., Özdamar, K., “Solving The Missing Value Problem By Use Of Simulated Data Sets”, ***Osmangazi Üniversitesi Tıp Fakültesi Dergisi***, **26**(2):67-76 (2004).

[3] Hedeker, D., Rose, J.S., “The natural history of smoking: A pattern-mixture random-effects regression model”, ***Multivariate Applications in Substance Use Research***, 79-112 (2000).

[4] Little, R.J.A., Rubin, D.R., “Statistical Analysis with Missing Data”, ***John Wiley&Sons***, New York (2002).

[5] Ibrahim, J.G., Molenberghs, G., **“**Missing data methods in longitudinal studies: a review”, ***Test,*** 18(1): 1-43 (2009).

[6] Allison, P.D., **“**Multiple imputation for missing data: a cautionary tale”, ***Sociological Methods and Research,*** 28,301–309 (2000).

[7] Yozgatlıgil, C., Aslan, S., İyigün, C., Batmaz, İ., Türkeş, M., Tatlı, H., “Comparison of methods to complete the missing data in time series: Turkey on the application of climate data”, ***Yöneylem Araştırması ve Endüstri Mühendisliği 30. Ulusal Kongresi***, 127- 128, 30 June - 2 July, Istanbul (2010).

[8] Yang, X., Shoptaw, S., “Assessing missing data assumptions in longitudinal studies: an example using a smoking cessation trial”, ***Drug and Alcohol Dependence,*** 77, 213-225 (2005).

[9] Baraldi, A.N., Enders, C.K., “An introduction to modern missing data analyses”, ***Journal of School Psychology,*** 48, 5-37 (2010).

[10]Yang, Y., Kang, J., “Joint analysis of mixed Poisson and continuous longitudinal data with nonignorable missing value”, ***Computational Statistics and Data Analysis,*** 54, 193-207 (2010).

[11]Fitzmaurice, G.M., Laird, N.M., Ware, J.H., “Applied Longitudinal Analysis”, ***John Wiley & Sons,*** New York (2004).

[12] Scheffer, J., “Dealing with missing data”, ***Res. Lett. Inf. Math. Sci.,*** 3, 153-160 (2002).

[13]Schaffer, J.L., Olsen, M.K., “Multiple imputation for multivariate missing-data problems: a data analyst’s perspective”, ***Multivariate Behavioral Research,*** 33(4): 545-571 (1998).

[14] SPSS, “IBM SPSS missing values 19” SPSS, Inc., ***IBM Company*** (2010).

[15]Ser, G., “Evaluation of the Multiple Imputation method regarding the quantitative characters with missing observations and covariance structures”, ***Journal of Animal and Veterinary Advances***, 10(24): 3269-3273 (2011).