Journal of Animal and Veterinary Advances 10 (24): 3269-3273, 2011
ISSN: 1680-5593

# Evaluation of the Multiple Imputation Method Regarding the Quantitative Characters with Missing Observations and Covariance Structures

Gazel SER
Department of Animal Science, Faculty of Agriculture, Yuzuncu Yil University, Van, Turkey

**Abstract:** The study aims to apply the Multiple Imputation (MI) method in case of missing observation in the quantitative data and to determine the covariance structure between the repeated measures. In estimating the missing observations, missing observations were assumed to be Missing at Random (MAR) and MCMC (Markov Chain Monte Carlo) technique and multiple imputation method were applied. To that end, live-weight data with missing observation and quantitative structure was used. Time factor was included as a continuous variable into the model that was formed to evaluate the live-weight data and random intercept and slope model were created. Compound Symetry (CS), Heterogenous Compound Symetry (CSH), Unstructured (UN), First order Autoregressive (AR (1)), Heterogenous first order Autoregressive (ARH (1)), Toeplitz (TOEP) and Heterogenous Toeplitz (TOEPH) structures were used to determine the covariance structure between repeated measurements in the data sets that have missing observations and missing observations which were estimated. Consequently, CS, AR (1) and TOEP structures were assumed to be the best model according to the AIC and BIC cohesion goodness of fit in modeling covariance matrix structure regarding the variable in the model established on the repeated measure data handled in both cases. UN, CSH, TOEPH and ARH (1) were found to be the worst model with heterogeneous covariance structure.

**Key words:** Repeated measures, markov chain monte carlo, goodness of fit

## INTRODUCTION

It is not always possible for all of the participating individuals or units to be in the study until the study is concluded. Lost or missing observation in the data set to be analyzed depends on either the conducting of study or material of the study conducted. It is essential to apply statistical methods to overcome missingness in the studies with such missing observations (Twisk, 2004; Baygul, 2007).

In the study conducted by Rubin in 1976, probability of missing observation in the data set are classified in three categories. Accordingly, missing observation Mechanism is 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 nonignorable if it depends on both observed and unobserved values (Nonignorable missing data or Missing Not At Random: MNAR) (Hedeker and Rose, 2000; Ibrahim and Molenberghs, 2009).

Multiple Imputation (MI) method is a statistical method which is commonly-used to estimate missing observation. MI method takes into account the MAR assumption for missing observation. The 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.

In the test designs including repeated measures, it is possible to get different features (Live-weight, height at withers, body length etc. in the field of animal) from variable structure test units with repeated measures, measures can be made in different times for same features (Tabachnick and Fidell, 2001). Repeated measures are acquired through carrying out several measures in the same test unit such as individual or animal at certain time intervals. The best example for such data in animal breeding is the data of growth curve in which each individual completes their growth at a specific time slice. The analysis of such measures acquired from the same test unit differs from classical analyses in some ways. Because there exists a relationship between the observations measured in the same unit. As known it is assumed that error terms in the analysis of variance are independent from each other and accordingly observations have equal correlation. In addition, observations which are close to each other in repeated measures are generally more correlated with the observations which are far from each others and have a heterogeneous variance. It is required to properly identify the covariance structure between the data in the analysis of the repeated measures.

One of the methods which take into the account missing observation situation in the data set and are used for modeling the variance-covariance structure in respect with the dependent variable is mixed models (Hedeker and Gibbson, 2006; Hogan, 2009).

In this research, random intercept and slope model in which time factor was included as a continuous variable were established on the data with repeated measure structure. The live-weight feature of the animals was handled as the dependent variable. Analyses were conducted at two stages. At the first stage, homogeneous and heterogeneous covariance structures were evaluated in order to determine the variance-covariance structure between the repeated measures in the data sets containing missing observations.

At the second stage in estimating the missing observations, missing observations were assumed to be Missing At Random (MAR) and the results pertaining to the same variance-covariance were acquired using MCMC technique and Multiple Imputation (MI) method. To that end, Compound Symmetry (CS), Heterogeneous Compound Symmetry (CSH), Unstructured (UN), First-Order Autoregressive (AR (1)), Heterogeneous First-Order Autoregressive (ARH (1)), Toeplitz (TOEP) and Heterogeneous Toeplitz (TOEPH) structures were utilized to determine the variance-covariance matrix structure of the dependent variable. In both cases, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used to determine the most coherent covariance for the model.

## MATERIALS AND METHODS

The 57 Il de France lambs whose weights were monitored at 14th day periodical intervals until the month of 6 starting from birth were used as animal material.

**Multiple imputation method:** In practice, missing observations in MI t.'ci iteration and calculation made in case of giving parameter estimation ($Y_{mis}^{(t)}$ and $\theta^{(t)}$) values are a MCMC process. Random variables are removed from the probability distributions by the help of Markov chains. The iteration steps of the method are accordingly as follows:

**Step 1:**

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

**Step 2:**

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

The step 1 is the imputation step while the step 2 is the posterior step. At the imputation step, sample estimate is made at random for the missing data from Pr ($Y_{mis}|Y_{obs}, \theta^{(t)}$) distribution. Such estimates are indicated as $Y_{mis}^{(t+1)}$. At the posterior step, $Y_{mis}^{(t+1)}$ values are placed and indicated parameter estimate indicated as $\theta^{(t+1)}$ is made with the Pr ($\theta \mid Y_{obs}, Y_{mis}^{(t+1)}$ probability distribution. Markov chain which consists of such estimates and converge to Pr ($Y_{mis}, \theta \mid Y_{obs}$) distribution is formed (Schaffer and Olsen, 1998; Little and Rubin, 2002; Sartori *et al.*, 2005; Kenward and Carpenter, 2009; Ser, 2011).

**General linear mixed models:** General linear mixed models are defined as follows:

$$Y = X\beta + Zu + e$$

Due to the fact that it is assumed that Y have a normal distribution and β regression parameter is same for all of the individuals in the equation, it takes place as the fixed effect in the model. u is the subject specifics regression coefficient and takes place as the independent (u~N (0, G)) and random effect in the model. X and Z are design matrices for the fixed and random effects. Presence of u in the model indicates the existence of heterogenity between the individuals at a lower level of the β regression coefficient. It is like (e~N (0, R)) and (R = Cov (e)) with e error vector. In brief when it is assumed that random effects follow a normal distribution (Kincaid, 2005):

$$E\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\text{Var}\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}$$

Also, conditional expected value of Y is E (Y|u) = Xβ+Zu. Marginal average of Y is E (Y) = Xβ (Fitzmaurice *et al.*, 2004; Venables and Dichmont, 2004).

## RESULTS AND DISCUSSION

The descriptive statistics of variables are shown in Table 1. Results of the multiple imputation method are shown in Table 2 and results of the cohesion results acquired from variance-covariance structures of data sets with missing observations and incomplete data are

Table 1: Descriptive statistics for the variables

| Variables | N | N[1] | Mean | Median | SD | Minimum | Maximum |
|---|---|---|---|---|---|---|---|
| Sex | 798 | - | 1.596 | 2.00 | 0.491 | 1.00 | 2.00 |
| Dam age | 798 | - | 2.737 | 2.00 | 1.345 | 1.00 | 7.00 |
| Birth type | 798 | - | 1.439 | 1.00 | 0.622 | 1.00 | 3.00 |
| Live weight | 769 | 29 | 19.720 | 20.80 | 9.292 | 3.20 | 46.50 |
| Live weight[2] | 798 | - | 19.687 | 20.65 | 9.285 | 3.20 | 46.50 |

[1]The number of missing observations; [2]The estimation results of multiple imputation

Table 2: Multiple imputation results

| | Live-weight | | | | | |
|---|---|---|---|---|---|---|
| Data | Imputation | N | Mean | SD | Minumun | Maximum |
| Original data | - | 769 | 19.7204 | 9.29172 | 3.2000 | 46.5000 |
| Imputed values | 1 | 29 | 19.9438 | 9.85337 | 2.1083 | 36.6930 |
| | 2 | 29 | 16.8901 | 10.19237 | 0.7445 | 34.7452 |
| | 3 | 29 | 19.3373 | 10.64286 | 0.6530 | 39.8634 |
| | 4 | 29 | 19.2370 | 9.82564 | 0.7090 | 35.8141 |
| | 5 | 29 | 18.8008 | 9.22717 | 3.8467 | 33.6868 |
| Complete data after imputation | 1 | 798 | 19.7285 | 9.30630 | 2.1083 | 46.5000 |
| | 2 | 798 | 19.6176 | 9.33409 | 0.7445 | 46.5000 |
| | 3 | 798 | 19.7065 | 9.33698 | 0.6530 | 46.5000 |
| | 4 | 798 | 19.7029 | 9.30562 | 0.7090 | 46.5000 |
| | 5 | 798 | 19.6870 | 9.28523 | 3.2000 | 46.5000 |

Table 3: The results of obtained from the goodness of fit incomplete and complete data sets of observed variance-covariance structures

| | Missing | | Complete | |
|---|---|---|---|---|
| Covariance structures | AIC | BIC | AIC | BIC |
| CS | 3939.4 | 3945.5 | 4139.9 | 4146.0 |
| CSH | 3939.5 | 3947.7 | 4140.9 | 4149.1 |
| AR (1) | 3939.4 | 3945.5 | 4139.9 | 4146.0 |
| ARH (1) | 3939.5 | 3947.7 | 4140.9 | 4149.1 |
| TOEP | 3939.2 | 3947.3 | 4140.3 | 4148.5 |
| TOEPH | 3939.5 | 3947.7 | 4140.9 | 4149.1 |
| UN | 3941.1 | 3951.4 | 4142.3 | 4152.5 |

Table 4: The results of fixed effects in the missing data set

| | Sex | | Dam age | | Birth type | | Time | |
|---|---|---|---|---|---|---|---|---|
| Covariance structures | F | p | F | p | F | p | F | p |
| CS | 46.51 | <0.0001 | 7.39 | 0.0067 | 11.40 | 0.0008 | 912.32 | <0.0001 |
| CSH | 39.46 | <0.0001 | 5.72 | 0.0170 | 8.03 | 0.0047 | 1022.96 | <0.0001 |
| AR (1) | 46.51 | <0.0001 | 7.39 | 0.0067 | 11.40 | 0.0008 | 912.37 | <0.0001 |
| ARH (1) | 39.46 | <0.0001 | 5.72 | 0.0170 | 8.03 | 0.0047 | 1022.96 | <0.0001 |
| TOEP | 44.42 | <0.0001 | 6.59 | 0.0105 | 8.71 | 0.0033 | 1039.54 | <0.0001 |
| TOEPH | 39.46 | <0.0001 | 5.72 | 0.0170 | 8.03 | 0.0047 | 1022.96 | <0.0001 |
| UN | 42.96 | <0.0001 | 6.30 | 0.0123 | 8.34 | 0.0040 | 1041.00 | <0.0001 |

Table 5: The results of fixed effects in the completed data set

| | Sex | | Dam age | | Birth type | | Time | |
|---|---|---|---|---|---|---|---|---|
| Covariance structures | F | p | F | p | F | p | F | p |
| CS | 51.07 | <0.0001 | 4.20 | 0.0409 | 13.22 | 0.0003 | 929.86 | <0.0001 |
| CSH | 45.55 | <0.0001 | 3.30 | 0.0697 | 10.58 | 0.0012 | 1016.29 | <0.0001 |
| AR (1) | 51.07 | <0.0001 | 4.19 | 0.0409 | 13.22 | 0.0003 | 929.76 | <0.0001 |
| ARH (1) | 45.55 | <0.0001 | 3.30 | 0.0697 | 10.58 | 0.0012 | 1016.29 | <0.0001 |
| TOEP | 49.21 | <0.0001 | 3.42 | 0.0648 | 10.79 | 0.0011 | 1039.92 | <0.0001 |
| TOEPH | 45.55 | <0.0001 | 3.30 | 0.0697 | 10.58 | 0.0012 | 1016.29 | <0.0001 |
| UN | 50.95 | <0.0001 | 3.58 | 0.0590 | 11.21 | 0.0009 | 1038.08 | <0.0001 |

shown in Table 3. Results of the fixed effects in the data sets with missing observations and incomplete data are in Table 4 and 5. In this research, random intercept and slope model in which time factor was modeled as a continuous variable in incomplete data with missing observations was utilized. Features apart from live-weight are discrete features. About 29 missing observations are available in the variable treated as the dependent variable (Table 1). Missing observation in the data set may be regarded as MAR in the event that such observations are <5%

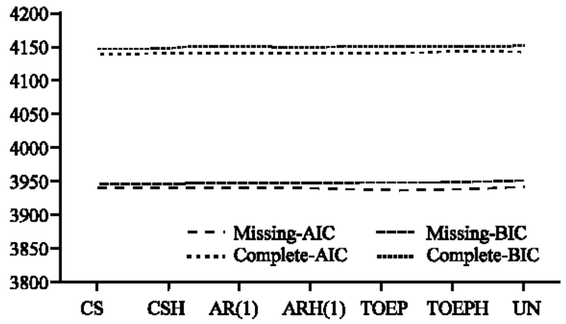

Fig. 1: Missing and complete data sets of observed values for the AIC and BIC fit criteria based on the covariance structure of exchange

(SPSS, 2010). Standard deviation, average, minimum and maximum values pertaining to each set of assigned values are shown in the table providing descriptive statistics for the live weight variable (Table 2).

In the MCMC method, parameter estimates are iteratively made with the probability distribution of the parameters of distribution and missing observations are randomly filled with these distribution estimates. MCMC method is utilized to ensure the estimated parameter to be independent from each other and to converge to a specific distribution. To this end, parameter results and cohesion measures were acquired by using the results of imputation in which the closest results to the original data are acquired at the assessment stage.

According to the Table 2 when the data acquired from imputation 5 are concerned standard deviation is closer to the original data and the minimum and maximum values are equal to the original data.

When the cohesion criteria are assessed (Table 3), homogeneous CS, TOEP and AR (1) structures were found to be the best model in the AIC and BIC goodness of fit in the repeated measures data with missing and incomplete observations which were handled in the modeling using homogeneous CS, TOEP, AR (1) variance-covariance matrix structure of the dependent variable in the random intercept and slope structure and UN, CSH, TOEPH and ARH (1) heterogeneous variance-covariance structures. UN, CSH, TOEPH and ARH (1) were found to be the worst model group with heterogeneous structure. Results of the cohesion criterion showed similar tendency in both cases according to the structures of variance-covariance applied in data set with incomplete and missing observation (Fig. 1).

## CONCLUSION

This study shows that it is essential to accurately model the variance-covariance so as for results of fixed effects to be applicable. Sex, dam age, birth type and time effects were found to be significant in each variance-covariance structure in the data set with missing observation (p<0.05). Some differences occurred between the F-values calculated. While the same results were acquired in respect with F-values in the heterogeneous variance-covariance structures apart from UN structure and same results were also acquired in homogeneous CS and AR (1) structures; different results were get from TOEP structure. While sex, birth type and time effects were found to be significant in each variance-covariance structure in the incomplete data set, dam age was only found to be significant in CS and AR (1) covariance structures (p<0.05). While the same results were acquired in respect with F-values in the heterogeneous variance-covariance structures and same results were also acquired in CS and AR (1) structures but different results were get from UN and TOEP structures.

Consequently, 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. Covariance structures between the measures should be also analyzed in the repeated measure data and probability of missing observation should be taken into consideration.

## REFERENCES

Baygul, A., 2007. Evaluation of the commonly used missing value analysis methods. M.Sc. Thesis, Istanbul University, Institute of Health Science, Biostatistics, Istanbul.

Fitzmaurice, G.M., N.M. Laird and J.H. Ware, 2004. Applied Longitudinal Analysis. 1st Edn., John Wiley and Sons Inc., New York, ISBN: 0-471-21487-6, pp: 187-234.

Hedeker, D. and J.S. Rose, 2000. The Natural History of Smoking: A Pattern-Mixture Random-Effects Regression Model. In: Multivariate Applications in Substance Use Research, Rose, J.S., L. Chassin, C.C. Presson and S.J. Sherman (Eds.). Lawrence Erlbaum, Hillsdale, New Jersey, pp: 79-112.

Hedeker, D. and R.C. Gibbson, 2006. Longitudinal Data Analysis. John Wiley and Sons, Inc., New Jersey.

Hogan, J.W., 2009. Comments on: Missing data methods in longitudinal data studies: A review. Test, 18: 59-64.

Ibrahim, J.G. and G. Molenberghs, 2009. Missing data methods in Longitudinal studies: A review. Test, 18: 1-43.

Kenward, M.G. and J.R. Carpenter, 2009. Multiple Imputation. In: Longitudinal Data Analysis, Fitzmaurice, G., M. Davidian, G. Verbeke and G. Molenberghs (Eds.). Taylor and Francis Group, CRC Press, New York, pp: 477-499.

Kincaid, C., 2005. Guidelines for selecting the covariance structure in mixed model analysis. Stat. Data Anal., 30: 1-8.

Little, R.J.A. and D.B. Rubin, 2002. Statistical Analysis with Missing Data. 2th Edn., John Wiley Publishers Company, New York.

SPSS, 2010. IBM SPSS missing values 19. SPSS, Inc., IBM Company.

Sartori, N., A. Salvan and K. Thomaseth, 2005. Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose. Comput. Stat. Data Anal., 49: 937-953.

Schaffer, J.L. and M.K. Olsen, 1998. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behav. Res., 33: 545-571.

Ser, G., 2011. Model selection and comparing optimization techniques in marginal and non-marginal multilevel generalized linear mixed model using missing observed longitudinal data. Ph.D. Thesis, Yuzuncu Yil University, Institute of Natural Science, VAN.

Tabachnick, B.G. and L.S. Fidell, 2001. Using Multivariate Statistics. 4th Edn., Allyn and Bacon, Boston, MA.

Twisk, J.W.R., 2004. Longitudinal data analysis. A comparison between generalized estimating equations and random coefficient analysis. Eur. J. Epidemiol., 19: 769-776.

Venables, W.N. and C.M. Dichmont, 2004. GLMs, GAMs and GLMMs: An overview of theory for applications in fisheries research. Fish. Res., 70: 319-337.