Sains Malaysiana 45(11)(2016): 1755–1761

# The Performance of Multiple Imputations for Different Number of Imputations

(Prestasi Pelbagai Imputasi untuk Bilangan Imputasi Berlainan)

GAZEL SER*, SIDDIK KESKIN & M. CAN YILMAZ

ABSTRACT

*Multiple imputation method is a widely used method in missing data analysis. The method consists of a three-stage process including imputation, analyzing and pooling. The number of imputations to be selected in the imputation step in the first stage is important. Hence, this study aimed to examine the performance of multiple imputation method at different numbers of imputations. Monotone missing data pattern was created in the study by deleting approximately 24% of the observations from the continuous result variable with complete data. At the first stage of the multiple imputation method, monotone regression imputation at different numbers of imputations (m=3, 5, 10 and 50) was performed. In the second stage, parameter estimations and their standard errors were obtained by applying general linear model to each of the complete data sets obtained. In the final stage, the obtained results were pooled and the effect of the numbers of imputations on parameter estimations and their standard errors were evaluated on the basis of these results. In conclusion, efficiency of parameter estimations at the number of imputation m=50 was determined as about 99%. Hence, at the determined missing observation rate, increase was determined in efficiency and performance of the multiple imputation method as the number of imputations increased.*

*Keywords: Multiple imputation; number of imputations; relative efficiency*

ABSTRAK

*Kaedah pelbagai imputasi adalah suatu kaedah yang digunakan secara meluas dalam menganalisis data yang hilang. Kaedah ini terdiri daripada proses tiga peringkat termasuk imputasi, analisis dan pengumpulan. Bilangan imputasi yang dipilih dalam langkah imputasi pada peringkat pertama adalah penting. Oleh yang demikian, kajian ini bertujuan untuk mengkaji prestasi pelbagai kaedah imputasi pada bilangan imputasi yang berbeza. Corak data hilang monoton telah dibentuk dalam kajian ini dengan menghapuskan kira-kira 24% pemerhatian daripada hasil berterusan pemboleh ubah dengan data yang lengkap. Pada peringkat pertama kaedah pelbagai imputasi, imputasi regresi monoton dalam bilangan imputasi yang berbeza (m=3, 5, 10 dan 50) telah dijalankan. Pada peringkat kedua, penganggar parameter dan ralat piawaian telah diperoleh dengan mengaplikasikan model linear umum kepada setiap set data lengkap yang diperoleh. Pada peringkat akhir, keputusan yang diperoleh telah dikumpulkan dan kesan bilangan imputasi ke atas penganggar parameter dan ralat piawai mereka dinilai berdasarkan keputusan ini. Kesimpulannya, kecekapan penganggar parameter kepada bilangan imputasi m=50 telah ditentukan sebanyak 99%. Oleh itu, pada kadar pemerhatian hilang yang ditentukan, kenaikan telah ditentukan dalam kecekapan dan prestasi kaedah pelbagai imputasi kerana jumlah imputasi meningkat.*

*Kata kunci: Jumlah imputasi; kecekapan relatif; pelbagai imputasi*

## INTRODUCTION

More continued interest on the analysis of missing observations has been received with the development of missing data mechanisms suggested by Rubin (1976), who divided missing observation mechanisms into three classes. The first of these mechanisms is structured MCAR (missing completely at random) that are independent from observed and unobserved data. The second is structured MAR (missing at random), depending only on the observed values. The last is structured MNAR (missing not at random), depending both on observed and unobserved values. Additionally, missing data structures shown by the missing observations in a data set are also important, as well as the missing data mechanisms, which is known as the missing data model. Missing data mechanism and missing data model have quite different meaning in missing data analysis. The missing data model do not provide information on why data is missing but only provide visual information on the distribution of missing observations in the data set. Missing data mechanisms, in contrast, mathematically deal with the relationship between observable data and missing observations, although no information is provided on the causes of occurrence of missing observation (Ender 2010). The missing data structures available for determining the nature of missing observation(s) in data set variables involve the unit nonresponse, monotone, planned missing, latent variable, general and univariate patterns. Formal illustrations of these models in a data set, as comprehensively described by Enders (2010), are provided in Figure 1. In all the models built by considering the four

variables given in Figure 1, the positions of the observable and missing values in the data set are provided visually. For instance, in the general model, we can see in the figure that missing observations are distributed randomly, rather than systematically. We note that in all models, we must determine whether or not a missing observation for one variable is dependent on another variable. The presence or absence of dependence between variables is defined by the missing data mechanisms. Conversely, for a missing observation for one variable in a data set in a single-variable structure, other variables cannot affect the missing observation (Enders 2010; Toka 2012).

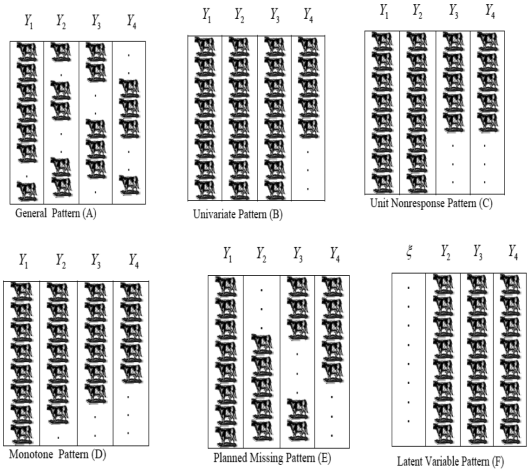



FIGURE 1. Missing data models (Adapted from Enders (2010))

In this study, we focus on a monotone model structure in which the number of missing observations in a data set increases with the number of variables. Considering the monotone model in Figure 1, whereas there are no missing observations in the first variable, the number of missing observations increases in the next three variables. The greatest number of missing observations occurs in the last variable. According to Enders (2010), the monotone model structure is frequently implemented in longitudinal studies. For example, participants may be omitted from a clinical study owing to their reaction to a new medicine tested. In addition, the order of missing observations provided by the monotone data model can decrease the confusion that arises in complex mathematical models, such as maximum likelihood (ML) and multiple imputations (MI).

The MI method specified for the monotone missing data model is a three-phase process involving imputation, statistical analysis and then pooling of the two. Proper values are imputed *m* times in lieu of the missing values, in respect to the number of imputations performed in the imputation step. In the second step, parameter estimations and standard errors are acquired by determined statistical analysis completed data sets obtained from the first step. In the last step, the results are interpreted by combining the obtained parameter estimations and standard errors through the formulas suggested by Harel (2007), Hippel (2005), Rubin (1987) and Schafer (1997). In the imputation phase, for the first step of the MI method, there are different approaches for determining the number of imputations to be used. For example, Hershberger and Fisher (2003) assert that hundreds of imputations are required; whereas Rubin (1987) claims that 2-10 imputations are adequate. Graham et al. (2007) and Schafer (1997) reported that better results may be obtained by as few as 3-5 imputations. Rubin and Schenker (1986) pointed out that the determinant in the selection of the number of imputations is the rate of missing observations in the data; if the rate is 10%, the number of imputation may be about 2 and if the rate is 60%, the number of imputations can just be 3. In addition, Hippel (2005) informed that hundreds of imputations are not necessary and that it takes approximately 2 h to process the data even when the number of imputations is approximately 20%, when implementing a complex model in a large data set having missing observations. When hundreds of imputations are used in a data set, several weeks and a vast computer memory storage area will be required. According to Graham (2012), while it is generally accepted by previous researchers that the imputation of a few data sets is adequate, the author advised that a greater number of imputations is required in recent studies and in particular, stated that this was required to achieve the statistical power equivalent to that of the ML procedure. Schafer (1997) indicated that better results are obtained with smaller numbers of imputations as 3-5 in MI for two reasons. First, MI is based on a simulation of the solution for only the missing data problem and the Monte Carlo method is used in this simulation. In any simulation technique in which Monte Carlo method is used, a selection of 'm' that is too big will efficiently eliminate any Monte Carlo error. However, as the Monte Carlo error will be insignificant in MI, the selection of a small number of imputations will result in acquiring more efficient results. Second, Schafer (1997) specified that the reason for the use of very few imputations as adequate is that the MI method combines the analysis results of the full data set, based on the number of imputations.

In this study, the monotone missing observation model is developed by deleting observations (approximately at a rate of 24%) of the dependent variables over the full data set and we then evaluate the performance of the MI method for different numbers of imputations (m = 3, 5, 10, 50).

## MATERIALS AND METHODS

### ANIMAL MATERIAL

We applied data on male lambs and kids in this study. The animals were exposed to two different feeding systems (concantrate and pasture) and their subsequent linolenic acid contents were measured in four different anatomic regions of animals after cutting (subcutaneous fat (SF), Longissimus dorsi (LD), Semimembranosus (SM)

and Triceps brachii (TB)). We took fatty acid measurements in the tissues of 47 animals aged 6-7 months.

### MULTIPLE IMPUTATION ANALYSIS

The analyses were performed using the PROC MI option in SAS (Version 9.4) packaged software (SAS 2014). First, we developed the monotone model data structure by randomly deleting observations in the full data set. Next, the analyses were performed in the following steps:

### MISSING VALUE IMPUTATION PHASE

In the first MI step, we determined the number of imputations (= nimpute) in the Proc MI procedure to be 3, 5, 10 and 50 and then transferred the completed data set, according to the numbers of imputations, to a new conclusion file 'a1.' In order to impute missing values in the fatty acid measurements, (linolenic acid content (lac)) were used the regression method. In the "class" case, the defined categorical variables were (male lambs and kids (spec), feeding system (fs), and anatomic regions (ar)). We note here that the 'class' case can only be used in a monotone missing data model (Berglund 2010).

```
proc mi data=a nimpute=3,5,10,50 out=a1 seed=54321;
class spec fs ar;
monotone regression;
var spec fs ar lac;
run ;
```

### STATISTICAL ANALYSIS PHASE

We obtained parameter estimations and standard errors by applying general linear model (GLM) analysis (Proc GLM) to the full data sets in the 'a1' file. On the basis of the number of imputations determined using the 'by_imputation' code, separate estimations are acquired for each imputation phase. The estimations obtained here were then restored in an output file 'a2.'

```
proc glm data=a1;
class spec   fs ar;
by _imputation_ ;
model lac=spec fs ar / ss3 solution ;
lsmeans spec fs ar/stderr;
means spec   fs ar/duncan;
ods output Parameter Estimates=a2 (where=(_imputation_ ne . ));
run;
```

### POOLING PHASE

In this last MI step, we combined the estimation results from the parameters treated by the mathematical equations derived by Rubin (1987) and then calculated the average of all individual estimations generated from the analysis of each imputed data set. $\hat{\theta}$ is the point estimation of the parameter ($\theta$) and the average of all point estimations of the parameters obtained from m pieces of completed data set may be written as follows:

$$\bar{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i. \tag{1}$$

The variance estimation obtained by the MI method has two components. First is the within-imputation variance, which shows the variability within each data set for which imputations are performed, calculated as follows:

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m}U_i. \tag{2}$$

Second is the between-imputation variance, which is the component showing the variation between the data sets for which imputation is performed, while also taking into account the uncertainty arising from the imputations. This is calculated mathematically as follows:

$$B = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\theta}_i - \bar{\theta}\right)^2. \tag{3}$$

The total variance comprising these two components is calculated as:

$$T = \bar{U} + (1 + m^{-1})B. \tag{4}$$

The standard error in MI is obtained with $\sqrt{T}$. Rubin (1987) calculated the fraction of the missing (FMI) value to estimate the rate of missing data in a population as follows:

$$FMI = \frac{(r+2)/(df+3)}{r+1} \text{ and } r = \frac{\left(1+m^{-1}\right)B}{U}. \tag{5}$$

FMI represents the amount of lost information arising from missing data when a parameter is estimated. In (5), the degrees of freedom (df) is calculated as follows:

$$df = (m-1)\left[1 + \frac{\bar{U}}{(B + B/m)}\right]^2.$$

Degrees of freedom in MI analysis is different from the df calculated using other statistical methods. It is not concerned with sample size. df is an indicator of the stability of the estimations. A small df indicates that the number of imputations is small and that the estimations acquired from this number of imputations are not stable. Likewise, when a df is considerably larger than the number of imputations, this indicates that the estimations obtained are stable and reliable (Enders 2010; Graham 2012; Schafer & Olsen 1998). The size of the standard errors, in the MI method, is determined using relative efficiency (RE):

$$RE = \left(1 + \frac{FMI}{m}\right)^{-1}. \tag{6}$$

As can be understood from (6), the efficiency of the MI method depends on the rate of the missing data in a population and the number of imputations carried out. According to Schafer and Olsen (1998), the number of imputations and the RE of the MI method are directly proportional, while the RE is inversely proportional to the rate of the missing data in the data set, i.e. the FMI. In

other words, as the number of imputations increases, the RE increases and as FMI increases, the RE decreases. The third combination step of the MI analysis is performed using PROC MIANALYZE code. Corrected parameter estimations and variance information are obtained in this combined transaction by using the equations above. In the case of 'Model effects', the variables accepted as reference parameters in the model are not written (kids (spec2), pasture (fs2) and TB).

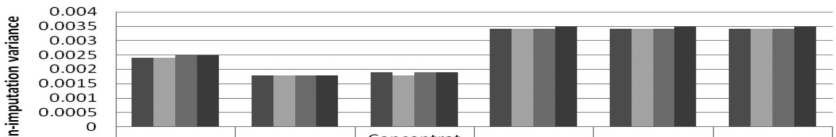
```
proc mianalyze parms=a2;
modeleffects intercept spec1 fs1 SF LD SM;
run;
```

## Results and Discussion

Variance information in the determined parameters, as the model effects are primarily shown with graphics (Figures 2-5) in the third step of the MI method. Specifically, within-imputation variance information is shown in Figure 2 and between-variance information is shown in Figure 3.

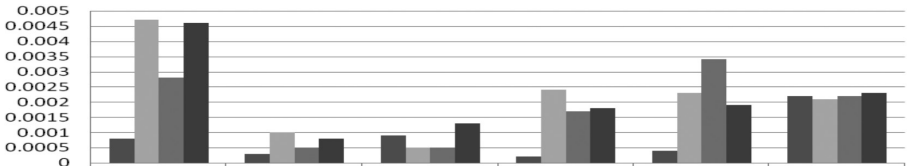
Within-imputations variance is the criterion that gives information about the variability within data sets that have no missing data (Enders 2010; Graham 2012). Variability in the intercepts, male lambs, concantrate and model effects in Figure 2 are fewer with respect to the SF, LD and SM. However, in general, the variability in the data sets for all number of imputations and for all parameters is nearly equal.

The between-imputation variance of the independent variables is the variability between the data sets, as shown in Figure 3, by taking the number of imputations into consideration. A review of the between-imputations variance results showed that while the variability between m=5 and the 50th imputation is greater for the intercept and the other parameters is smaller. Between-imputations variability is the criterion for taking into account the variability between parameter estimations calculated from the imputed data sets. When the variability among estimations is large, the variability arising from the number of imputations is high and therefore the between-imputations variance is also large (Enders 2010; Wayman



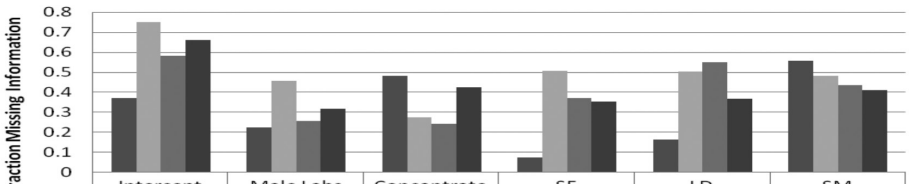

| | Intercept | Male Labs | Concantrate | SF | LD | SM |
|---|---|---|---|---|---|---|
| m=3 | 0.0024 | 0.0018 | 0.0019 | 0.0034 | 0.0034 | 0.0034 |
| m=5 | 0.0024 | 0.0018 | 0.0018 | 0.0034 | 0.0034 | 0.0034 |
| m=10 | 0.0025 | 0.0018 | 0.0019 | 0.0034 | 0.0034 | 0.0034 |
| m=50 | 0.0025 | 0.0018 | 0.0019 | 0.0035 | 0.0035 | 0.0035 |

FIGURE 2. Results of within-imputation




| | Intercept | Male Labs | Concantrate | SF | LD | SM |
|---|---|---|---|---|---|---|
| m=3 | 0.0008 | 0.0003 | 0.0009 | 0.0002 | 0.0004 | 0.0022 |
| m=5 | 0.0047 | 0.001 | 0.0005 | 0.0024 | 0.0023 | 0.0021 |
| m=10 | 0.0028 | 0.0005 | 0.0005 | 0.0017 | 0.0034 | 0.0022 |
| m=50 | 0.0046 | 0.0008 | 0.0013 | 0.0018 | 0.0019 | 0.0023 |

FIGURE 3. Results of between-imputation




| | Intercept | Male Labs | Concantrate | SF | LD | SM |
|---|---|---|---|---|---|---|
| m=3 | 0.3718 | 0.2234 | 0.4825 | 0.0759 | 0.1652 | 0.557 |
| m=5 | 0.7506 | 0.4566 | 0.2765 | 0.5065 | 0.5028 | 0.4809 |
| m=10 | 0.5834 | 0.2583 | 0.2442 | 0.3713 | 0.5493 | 0.4353 |
| m=50 | 0.6606 | 0.3189 | 0.4238 | 0.3529 | 0.3682 | 0.4109 |

FIGURE 4. Results of fraction of missing information (FMI)

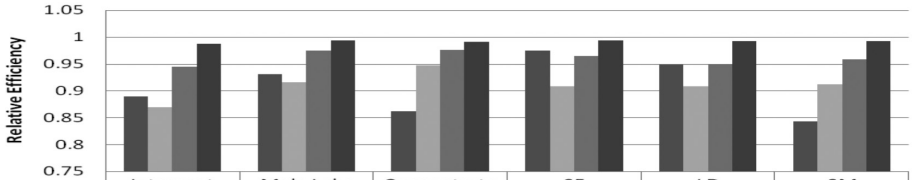
2003). FMI and RE values are provided in Figures 4 and 5, respectively.

FMI values with respect to the number of imputations and each parameter are given in Figure 4. FMI is defined in several sources as simply the missing data percentage. However, in contrast to the known percentage, FMI values are obtained after the correction is made by considering the high correlation between variables having missing data and other variables. FMI values are directly related to the number of imputations (Graham 2012). More stable values are obtained with greater numbers of imputations. In Figure 4, we see that the FMI value is not stable even for m=50 imputations. However, it is significant that it is more stable than for m=3 imputations. Enders (2010) reported that FMI is a criterion equivalent to the coefficient of determination ($R^2$) in MI and may be defined as the amount of variability in the variance of a parameter explained by missing data.

Figure 5 illustrates the relative efficiency of the MI method, as calculated with respect to FMI and the number of imputations for each parameter within the imputed data set. Also seen in the figure, as m increases, the efficiency of the MI method increases. Accordingly, for m=50 number of imputations, efficiency in all of the parameter estimations obtained in the study is approximately 99%. Therefore, the greater the number of imputations, the greater the efficiency and the better performance of the MI method as the rate of missing observations is determined. Schafer and Olsen (1998) reported that, for a high proportion of 30% in missing observation (FMI=0.3) used widely in application, the number of imputations m=5 and m=10 produced very high proportions of 94% and 97% in MI efficiency, respectively; therefore, the increment in imputation number provided nearly the same advantage in MI efficiency. Figure 6 depicts the obtained results for the FMI and relative efficiency values for each parameter according to the number of imputations.

In Figure 6, the lower the rate of missing data in a data group (FMI), the greater is the relative efficiency. For example, at m=3 imputations, RE increases while FMI decreases. In addition, at m=50 imputations, relative efficiency reaches 99% while FMI decreases. The efficiency of the MI method is directly proportional to the number of imputations and inversely proportional to FMI. Combined parameter estimations and standard errors obtained from m=50 imputations are shown in Table 1.

In Table 1, slightly bigger dfs are obtained for m=50 imputations. Thus, we may conclude that the estimations here are more reliable than those of other imputation numbers. According to Graham (2012), parameter estimations, standard errors, t (*Estimate*/*Std.Error*) and p values are keys to be used in the evaluation of MI estimations. Also, df and FMI (Figure 4) measurements are two important parameters that have a role in determining the number of imputations in the MI method. In addition, when df is deemed to be an indicator of stability and reliability of the obtained estimations, if the df is significantly greater than the number of imputations (Table 1), then the MI estimations may be assumed to be stable and reliable. However, if the df is small, the estimations from MI may be assumed to be stable and a greater number of imputations is required (Graham 2012). In the previous study conducted by Schafer and Olsen (1998), they stated that if the df is too small, the number of imputations must be increased in order to achieve more reliable and correct estimations.




| | Intercept | Male Labs | Concantrate | SF | LD | SM |
|---|---|---|---|---|---|---|
| m=3 | 0.8897 | 0.9306 | 0.8614 | 0.9753 | 0.9477 | 0.8433 |
| m=5 | 0.8694 | 0.9163 | 0.9475 | 0.908 | 0.9086 | 0.9122 |
| m=10 | 0.9448 | 0.9748 | 0.9761 | 0.9641 | 0.9479 | 0.9582 |
| m=50 | 0.9869 | 0.9936 | 0.9915 | 0.9929 | 0.9926 | 0.9918 |

FIGURE 5. Results of relative efficiency (RE)

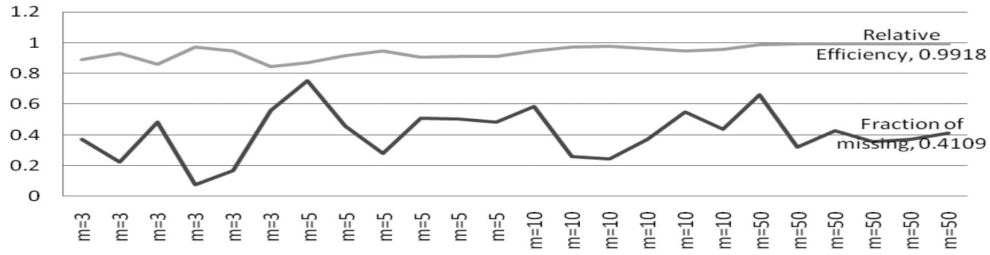



FIGURE 6. Results for relative efficiency (RE) and fraction of missing information (FMI)

TABLE 1. Results of pooling parameter estimations and standard errors obtained from m=50 imputations

| Parameter | Estimate | Standart Error | 95% Confidence Limits | | df | t | P |
|---|---|---|---|---|---|---|---|
| Intercept | 1.826829 | 0.085358 | 1.65774 | 1.99592 | 114.3 | 21.40 | <.0001 |
| Male lambs | 0.485378 | 0.052652 | 0.38193 | 0.58883 | 490.16 | 9.22 | <.0001 |
| Concantrate | -1.388232 | 0.057475 | -1.50137 | -1.27509 | 278.13 | -24.15 | <.0001 |
| SF | -0.553227 | 0.073432 | -0.69759 | -0.40887 | 400.61 | -7.53 | <.0001 |
| LD | -0.153776 | 0.074302 | -0.29989 | -0.00767 | 368.08 | -2.07 | 0.0392 |
| SM | 0.040741 | 0.076896 | -0.11059 | 0.19207 | 295.82 | 0.53 | 0.5966 |

With the MI method, different approaches are available for determining the optimal number of imputations. Some of these approaches suggest that small numbers of imputations (for example, 3-10 imputations) are sufficient and others state that hundreds of imputations are required. In studies with large data sets and when complex models are being implemented, the problem with MI analysis is that carrying out hundreds of imputations will take an unreasonable amount of time. The data set used in this study is small and the model established by the MI method is a simple GLM model. Nevertheless, while the GLM procedure operates at 3.79 s (real time) for m = 3 imputations, the results for the GLM procedure require 1:13.93 min (real time) for m = 50 imputations. Therefore, we see that there are probably serious time problems with complex models and large data sets.

## CONCLUSION

For m=50 imputations, our model results obtained 99% for all parameters including MI efficiency. Specifically, we obtained efficient estimations according to m=3 and 5 imputations. Based on these results, we can draw the following important conclusions with regards to the use of the MI: If the data set is not small and the statistical model to be implemented is not too complex, greater numbers of imputations may be preferable. When MI is applied to small data sets having a monotone missing data model, in which the dependent variables have constant and categorical covariates, valid results may be obtained using the GLM statistical model. Furthermore, no serious problems will occur with respect to time if a greater number of imputations is used, since the GLM model is not complex. MI provides efficient results at m=50 imputations under the above conditions.

## ACKNOWLEDGEMENTS

We appreciate Prof. Dr. Aşkın KOR and Assist. Prof. Dr. Serhat KARACA for their contributions on providing the data used as material of this study. An abstract of this paper was published as a book of abstracts in The 8th Conference of Eastern Mediterranean Region of the International Biometric Society (EMR 2015), May 11-15 2015, in Cappadocia, Nevşehir, Turkey.

## REFERENCES

Berglund, P.A. 2010. An introduction to multiple imputation of complex sample data using SAS v9.2. In *SAS Global Forum*. pp. 1-12.

Enders, C.K. 2010. *Applied Missing Data Analysis*. New York: The Guilford Publication.

Graham, J.W. 2012. *Missing Data: Analysis and Design*. New York: Springer Sciences & Business Media.

Graham, J.W., Olchowski, A.E. & Gilreath, T.D. 2007. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science* 8(3): 206-213.

Harel, O. 2007. Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology* 4(1): 75-89.

Hershberger, S.L. & Fisher, D.G. 2003. A note on determining the number of imputations for missing data. *Structural Equation Modeling* 10(4): 648-650.

Hippel, P.T. 2005. Teacher's corner: How many imputations are needed? A comment on Hershberger and Fisher (2003). *Structural Equation Modelling* 12(2): 334-335.

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

Rubin, D.B. 1976. Inference and missing data. *Biometrika* 63(3): 581-592.

Rubin, D.B. & Schenker, N. 1986. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* 81(394): 366-374.

SAS. 2014. *SAS/STAT, Statistical Analysis System for Windows. Relase 9.4*. Cary, NC, USA: SAS Institute Inc.

Schafer, J.L. & Olsen, M.K. 1998. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research* 33(4): 545-571.

Schafer, J.L. 1997. *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall/CRC.

Toka, O. 2012. Robust estimation in case of missing data. MSc diss., University of Hacettepe, Ankara, Turkey (Unpublished).

Wayman, J.C. 2003. Multiple imputation for missing data: What is it and how can I use it? In *Annual Meeting of the American Educational Research Association,* Chicago, IL. pp. 2-16.

Gazel Ser* & M. Can Yılmaz
Department of Animal Science, Biometry and Genetics Unit
Faculty of Agriculture, University of Yuzuncu Yil
65080 Van
Turkey

Sıddık Keskin
Department of Biostatistics, Faculty of Medicine
University of Yuzuncu Yil
65080 Van
Turkey

*Corresponding author; email: gazelser@gmail.com

Received: 17 September 2015
Accepted: 17 March 2016