HACETTEPE UNIVERSITY, INSTITUTE OF POPULATION STUDIES
TECHNICAL DEMOGRAPHY PROGRAM

AN ANALYSIS OF A MULTIPLE IMPUTATION MODEL FOR THE MISSING VALUES IN SELECTED VARIABLES OF TDHS-2003 DATA: THE CASE OF ANTHROPOMETRIC MEASURES

Bengi UĞUZ

M.A. thesis submitted for the partial fulfilment
of the requirements for the M.A. degree
in Technical Demography Program at Hacettepe University
Institute of Population Studies

Ankara, June 2007

HACETTEPE UNIVERSITY, INSTITUTE OF POPULATION STUDIES
TECHNICAL DEMOGRAPHY PROGRAM

AN ANALYSIS OF A MULTIPLE IMPUTATION MODEL FOR THE MISSING VALUES IN SELECTED VARIABLES OF TDHS-2003 DATA: THE CASE OF ANTHROPOMETRIC MEASURES

Bengi UĞUZ

M.A. thesis submitted for the partial fulfilment
of the requirements for the M.A. degree
in Technical Demography Program at Hacettepe University
Institute of Population Studies

Supervisor

Dr. Ahmet Sinan TÜRKYILMAZ

Ankara, June 2007

ACCEPTANCE AND APPROVAL

This is to certify that we have read and examined this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Arts in Technical Demography.

Examining Committee Members:

Member (Supervisor):……
Dr. Ahmet Sinan Türkyılmaz, Hacettepe University, Institute of Population Studies

Member (Chair):……
Assoc.Prof.Dr. İsmet Koç, Hacettepe University, Institute of Population Studies

Member:……
Assoc.Prof.Dr. Turgay Ünalan, Hacettepe University, Institute of Population Studies

This thesis has been accepted by the above-signed members of the Committee and has been confirmed by the Administrative Board of the Institute of Population Studies, Hacettepe University.

Date: / / 2007

Prof. Dr. Sabahat Tezcan
Director

ABSTRACT

Sample surveys are one of the three main sources of social and demographic data together with the population censuses and vital registrations. They are also the most important source for the countries that are lacking well-established registration systems, such as Turkey. The issue of missing data is a common problem in social surveys and cause biased estimations if not dealt properly. However, the number of theoretical and practical studies on techniques for handling missing data is very limited in Turkey.

Demographic and Health Surveys (DHS), on the other hand, use some well-established editing and imputation techniques for only dates for several key events. Being one of the most widespread surveys of the world, many variables are exposed to missing data and inconsistency problems in DHS, although its complex design and questionnaire structure aims at minimizing these problems. In addition, existing imputation techniques of DHS have their own shortcomings.

The overall objective of this study is to apply the multiple imputation model to the anthropometric measurements of children under age five, in the 2003 Turkey Demographic and Health Survey (TDHS–2003) data, which is implemented by the Hacettepe University Institute of Population Studies. More specifically, sequential regression multiple imputation technique is used for creating 20 completed data sets, which are imputed conditional on the fully observed variables. The completed data sets are then analyzed and the results are compared with the observed data set, particularly for the anthropometric indexes height for age, weight for height and weight for age.

According to the results of the study, multiply imputed data well imitated the observed data in terms of distribution for both of the study variables, weight and height. The percentages below certain levels of anthropometric indexes slightly decreased after the imputation application, which indicate a possible bias among the children who were measured and not measured. Moreover, several analyses showed that the percentage of missing data differentiate substantially among some background characteristic categories of the respondent. As a result, multiple imputation corrected for the bias due to nonresponse in weight and height variables and increased reliability in parallel with the increase in number of observed cases.

ÖZET

Örneklem araştırmaları, nüfus sayımları ve kayıtlarla birlikte en önemli üç sosyal ve demografik veri kaynağından birini oluşturmaktadır. Örneklem araştırmaları, özellikle Türkiye gibi kayıt sistemi yerleşik olmayan ülkeler için en önemli veri kaynağını oluşturmaktadır. Kayıp veri konusu sosyal araştırmalarda oldukça sık rastlanılan bir konudur ve uygun bir şekilde ele alınmadığında yanlı tahminlere neden olmaktadır. Buna rağmen, Türkiye'de kayıp veri konusuyla ilgili teorik ve uygulamalı çalışmaların sayısı oldukça kısıtlıdır.

Diğer taraftan Nüfus ve Sağlık Araştırmaları (NSA), yalnızca bazı önemli tarihler için kullanılan bir imputasyon yöntemine sahiptir. Dünyanın en yaygın araştırmalarından biri olan NSA'da, araştırmanın karmaşık yapısı ve soru kağıdı tasarımı bunu en aza indirmeye çalışsa da, birçok değişken kayıp veri ve tutarsızlık sorunuyla karşı karşıyadır. Ayrıca var olan imputasyon tekniklerinde de çeşitli sorunlar bulunmaktadır.

Bu çalışmanın genel amacı, Hacettepe Üniversitesi Nüfus Etütleri Enstitüsü tarafından gerçekleştirilen 2003 Türkiye Nüfus ve Sağlık Araştırması'nda yer alan beş yaş altı çocukların antropometrik ölçümleri için çoklu imputasyon modelinin denenmesidir. Çalışmada daha özelinde, sıralı regresyon çoklu imputasyon tekniği kullanılmış ve 20 tamamlanmış veri seti elde edilmiştir. Bu veri setleri daha sonra analiz edilmiş ve gözlenmiş olan veri seti sonuçlarıyla, antropometrik ölçümler – yaşa göre boy, boya göre kilo ve yaşa göre kilo – bağlamında karşılaştırılmıştır.

Çalışmanın sonuçlarına göre, çoklu imputasyon tekniği ile elde edilen veri gözlenmiş olan veriyi, her iki çalışma değişkeni olan boy ve kilo için dağılım anlamında iyi taklit etmektedir. İmputasyon uygulamasından sonra, antropometrik endeksler için belli değerlerin altında kalan çocukların yüzdesi bir miktar düşmüştür, bu da ölçümü yapılmış ve yapılmamış çocuklar arasında muhtemel bir yanlılığa işaret etmektedir. Ayrıca, çeşitli analizler kayıp veri yüzdelerinin, cevaplayıcının belli özelliklerine göre değiştiğini ortaya koymaktadır. Çalışmada sonuç olarak çoklu imputasyon, boy ve kilo değişkenlerinde var olan cevapsızlığa bağlı yanlılığı düzeltmiş ve gözlem sayısındaki artışa paralel olarak güvenirliği artırmıştır.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Sample surveys are one of the three main sources of social and demographic data together with the population censuses and vital registrations. They are also the most important source for the countries that are lacking well-established registration systems, such as Turkey. The issue of missing data is a common problem in social surveys. Missing data over certain levels can cause biased estimations and defective interpretations if not dealt properly. Even though certain precautions are taken at design and pre-field stages of the survey, missing data problem often cannot be disposed fully and require editing and imputation practices at analysis stage. The overall objective of this study is to apply the multiple imputation model for the anthropometric measurements of children in TDHS–2003 data. Major motivation for the selection of this subject is the deficiency of methods for handling missing data problem in the analysis of social surveys, which affect the development plans of Turkey intimately and used in comparisons with the other countries. There are only a limited number of theoretical studies on the issue as well. Turkish Statistical Institute (TURKSTAT), which is the responsible institution for producing statistics in Turkey, also doesn't have structured imputation schemes and uses ad-hoc methods.

The existence of the 2003 Turkish Demographic and Health Survey (TDHS–2003) is seen as a noteworthy advantage in the investigation of the content of this study. Demographic and Health Surveys (DHS) are implemented to provide data on population, health and nutrition of women and children in developing countries of the world. TDHS–2003 is the third and latest available survey in Turkey, implemented by employing DHS methodology. Being one of the most widespread surveys of the world, many variables are exposed to missing data and inconsistency problems in DHS, although its complex design and questionnaire structure aims at minimizing these problems. DHS uses some well-established editing and imputation techniques for only dates for several key events. Missing values and inconsistencies in all other variables are flagged but not imputed. Croft (1991) discusses several

problem areas emerge at the imputation process of DHS surveys, and concludes as "… as techniques for survey data collection and for data processing have improved, so has the quality of data produced. The need for data editing and imputation techniques, serves to indicate that there is still a long way to go." (Croft, 1991). Therefore, it is believed that there is a need for debate for more robust imputation procedures in DHS.

Another issue composing the rationale for the study is related with the study variables selected. As is known, anthropometric measures of children under age five, which are mainly weight and height, are used in constructing indicators such as stunting, wasting and underweight. These indicators are used in monitoring and evaluation purposes for the development progress of countries. "The prevalence of underweight children" is the key indicator for the first goal of the UN's Millennium Development Goals which is an important treaty at international agenda (www.un.org/milleniumgoals/). Anthropometric measurement variables contain serious proportions of missing values in all of the three TDHS conducted in Turkey. Furthermore, it is a known fact that there are significant disparities among geographic regions of Turkey in terms of socio-demographic characteristics, which may cause problems for specific analyses when there are high levels of missing data for specific regions. As an example for that, it has been seen that approximately 20% of the data is missing for some date variables at country level increases to 40% for some regions at regional level, when the TDHS–2003 raw data[1] is examined. In a similar manner, anthropometric measurements for children display relatively higher proportions of missing data, at percentages over 10% for some regions.

There are different techniques to handle missing data problem in the literature from doing nothing to completely making up the data, which change in respect of the percentage of missing data, mechanism of missingness as well as the purpose of the analyses. In this study, multiple imputation technique will be used in order to handle

[1] Before DHS editing and imputation procedure is applied.

the missing data problem in anthropometric measurements in TDHS-2003 data. Being one of the most advanced imputation techniques, multiple imputation basically refers to "fill-in" the missing data by drawing from conditional distribution of the missing data given the observed data (Rubin, 1988). As its name specifies, more than one complete data set are created by multiple imputation, and analysis is conducted by considering the natural variability as well as the additional uncertainty created by imputation. In this thesis, twenty sets of imputed data are created to allow the assessment of variability due to imputation. The imputation procedure incorporated many predictors, including demographic and health-related variables. Four types of auxiliary variables are included to the imputation model, to which the imputations would be conditioned. These types of auxiliary variables contain (a) the essential information about the sample design, (b) demographic variables, (c) the ones preserving important statistical relationships between variables, and (d) variables related to the missingness. The significant variables are included according to the results of a stepwise regression analysis. Twenty sets of complete data are then analyzed according to the specified formula. Finally, results of multiple imputation are compared with the TDHS-2003 results.

In this context, the purpose of this study is to apply multiple imputation technique for the missing values in anthropometric variables in TDHS-2003 data, in order to obtain more reliable and valid results with more observations. The study is an attempt to provide new insight to the survey analysis practices in terms of handling missing data in Turkey, and it is expected to contribute to the DHS literature regarding missing data problem as well.

The study is organized as follows. Chapter 2 presents the conceptual framework on missing data, including a general overview as well as definitions and terminology for missing data and imputation. Types of imputation including the relevant literature review are also introduced in this chapter. Then, previous applications of imputations to large data sets and types of imputation methods implemented in DHS surveys are

presented and imputation related studies in Turkey are presented by the end of this chapter.

In Chapter 3, firstly the data source used in the study, namely the TDHS-2003 is introduced, with an additional section to examine the missing data in the data in general terms. Anthropometric variables and calculation of anthropometric indexes are given in this chapter, and the missing data on anthropometric measurements in TDHS-2003 data is discussed. In the introduction of estimation techniques, multiple imputation with a special emphasis to the selected imputation technique, namely sequential regression multivariate imputation and the software used are provided. Finally study data file and the study variables are explained, as well as the assumptions for multiple imputation. In this part, construction of the imputation model of the study is also presented.

In Chapter 4, results of multiple imputation in terms of its influence on the distribution of the variables of interest is presented firstly. Results of the applied multiple imputation are given in comparison with the observed data for the study variables in relation with some background characteristics as well. The results are concluded and discussed in the last chapter, namely Chapter 5. Some possible further studies are also mentioned in this chapter.

2. CONCEPTUAL FRAMEWORK

2.1. General Overview

Standard statistical methods usually consider that data sets have a rectangular form, where rows represent units and columns represent variables. However, in real world data sets values are generally incomplete in either rows or columns, since information cannot be collected completely from all elements in selected sample. A differentiation is made between unit nonresponse; when none of the survey responses are available for a sampled element; and item nonresponse; when some of the responses are available.

Unit nonresponse occurs because selected sample elements are unable to be contacted or refuse to participate in survey. In some cases, a certain level of unit nonresponse is taken into account and incorporated into sample size at design stage, in the light of previous experiences. Hence, no additional editing is performed. In other cases, the only information available about unit nonresponse is on sampling frame from which the sample was selected. Therefore, weighting adjustments are used to compensate unit nonresponse, where the sample size is artificially enlarged to reach the original sample size. In both cases, it is assumed that results are not biased between the responding and non-responding sample members.

Item nonresponse arises because of item refusals, “don’t know”s, omissions and answers deleted in editing. In the existence of item nonresponse, there are additional information (i.e. other responses) available for the elements involved. Methods used to handle item nonresponse are generally investigated under the general heading of imputation. The basic idea of imputation is to fill each missing datum with

reasonable guesses and conduct the analysis as if there were no missing data (Allison, 2001).

There are some serious problems created by missing data in a survey. First, the participants not included in the analysis may have different characteristics from those who were included. This can lead the analysis be biased. Second, the existence of missing data often implies a loss of information which makes the sample less representative and estimates less efficient than planned. Finally, standard software as well as statistical methods is designed for complete data sets. Researchers have to drop some units from the analysis because of incomplete data in certain cases. Hence, missing data may influence both the analysis and interpretation of the data.

2.2. Definitions and Terminology

2.2.1. Missing Data Mechanisms

All missing data strategies hold assumptions about the nature of the mechanism that causes the missing data. Missing data mechanisms are commonly described in three categories, described by Little and Rubin (1987):

- First, data can be "Missing Completely at Random" (MCAR). When data are MCAR, complete cases are a random sample of the originally identified set of cases. Since the complete cases are representative of the originally identified sample, inferences based on only complete cases are applicable to the target population.

- Second, data can be missing "Missing at Random" (MAR). In this case, missing data depends on known values and thus is described fully by variables observed in the data set. When data are MCAR or MAR, the response mechanism is termed to be *ignorable*. Ignorable response

mechanisms are important because when they exist, a researcher can ignore the reasons for missing data in the analysis of the data and simplify the model-based methods used for missing data analysis (Pigott, 2001).

- Third, data can be missing "Missing Not at Random" (MNAR) or "Not Missing at Random" (NMAR). This case is called *nonignorable*. With nonignorable missing data, the reasons for the missing observations depend on the values of those variables. Allison (2001) states that robust prior knowledge is required for effective estimation with nonignorable nonresponse, since the data contain no information about what models would be appropriate. Little and Rubin (1987) and Schafer (1997) discuss methods that can be used for non-ignorable missing data.

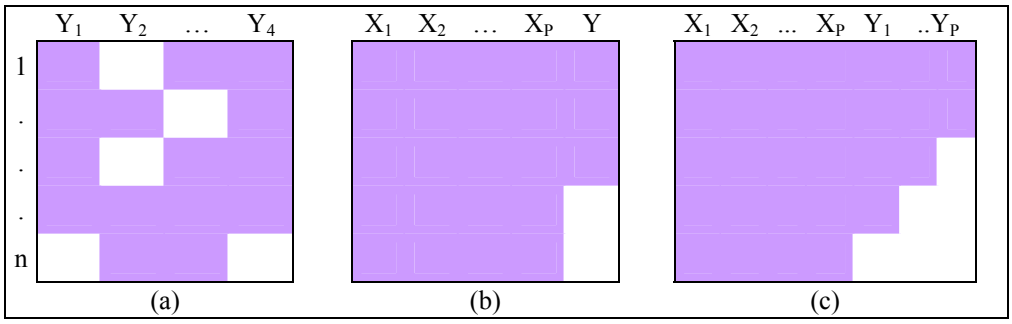
During data collection, the researcher has the opportunity to observe the possible explanations for missing data; evidence that will help guide the decision about what missing data method is appropriate for the analysis. However, it is generally impossible to measure the mechanism that holds for missing data. The data is assumed to be MAR or ignorable in most of the imputation procedures (Pigott, 2001).

2.2.2. Missing Data Patterns

In general, missing data pattern can be *univariate***,** which means that missing data values only occur in a single response variable, or *multivariate* in the sense that missing values occur in more than one variable. The general multivariate pattern is illustrated in Figure 2.1 (a) below. It can be seen that any set of variables may be missing in general multivariate pattern. For univariate case, which is shown in Figure 2.1 (b), the missing values occur on an item Y but a set of p other items X_1, $X_{2,\ldots,}X_p$ is completely observed. A particular missing data pattern is a *monotone* nonresponse, which may arise by dropouts in longitudinal studies. In Figure 2.1 (c), items or item

groups $Y_{1,\ldots,}Y_Q$ may be ordered in such a way that if Y_j is missing for a unit, then $Y_{j+1},..,Y_Q$ are missing as well; which shows a monotone pattern. The choice of imputation method generally depends on the underlying missing data pattern since the investigation of the nonresponse pattern is important and useful (Durrant, 2005).

Figure 2.1 Patterns of nonresponse in rectangular data sets: (a) general multivariate pattern, and special cases, (b) univariate pattern, and (c) monotone pattern



Source: Durrant, 2005.

2.3. Types of Imputation Techniques

Generally there are two ways to deal with missing values. First one is deletion which includes listwise deletion and pairwise deletion. In deletion, missing cases or items are precluded from the analysis. The second way is imputation which includes single imputation and multiple imputation.

Schafer and Graham (2002) stated that until the 1970s missing values were handled primarily by editing. The principal source on incomplete data is written by Rubin (1976), which is still an important reference today. With the improvements in methodologies and software facilities, many techniques to deal with incomplete data were developed and employed in different disciplines. Later, Little and Rubin (1987) presented a very detailed review of current methodologies on a theoretical basis. In the following, a review of available imputation methods is given considering the historical sequence, in terms of classical methods and multiple imputation.

2.3.1. Classical Methods

Before 1980s, most widely applied methods to deal with missing data were ad-hoc, including deletion methods and single imputation methods. *Listwise deletion*, also known as complete-case analysis or case deletion is accomplished by deleting incomplete units from the sample. It is used by default in many statistical programs. There are two advantages of listwise deletion according to Allison (2002); (1) it can be used for any kind of statistical analysis, and (2) special computational methods are not required. He adds that, when data are MCAR, the reduced sample will be a random sub-sample of the original sample (Allison, 2002). However, Pigott (2001) pointed out that even when data are MCAR, listwise deletion has some potential difficulties. In particular, in large datasets there may be few cases with all variables observed. If incomplete cases are simply discarded, few units may be left for the analysis. Moreover, information contained in incomplete data is ignored by deletion, where it could provide some insight about the outcome of interest. Because as Little and Rubin (1987) stated, completely recorded units usually differ in important ways from the original sample.

In *pairwise deletion*, also called available case analysis, summary statistics to be used for linear models are computed using all units that are available to compute that one statistic. Allison (2002) provides an instance such that; to compute the covariance between two variables X and Z, all units that have data present for X and Z are used. Pairwise deletion makes use of a considerable amount of data compared to listwise deletion. Schafer and Graham (2002) emphasized that the samples used to estimate parameters are different from each other, thus it is difficult to compute standard errors or other measures of uncertainty, which lead the analysis be problematic.

Imputation, the practice of filling in the missing data, has both its advantages and disadvantages. Schafer and Graham (2002) summarized the general features of imputation as follows: Imputation is more efficient than case deletion, because no units are excluded and the original sample size is preserved. As well as that, imputation makes use of information covered by missing values that is destroyed by deletion methods. Imputation also produces an apparently complete data set that may be analyzed by standard methods and software. Last, the nonresponse problem is solved in the same way for all users of the dataset once, thus the analyses will be consistent across all users. While imputation allows the inclusion of all units in a standard analysis procedure, replacing missing values with a single value changes the distribution of that variable by decreasing the variance that is likely present (Pigott, 2001).

Mean/median imputation is one of the most frequently used imputation methods. Missing values are replaced by the average or the median of the observed values for that item. By this procedure the average of the variable is preserved but variance and covariance terms will be negatively biased because any missing observation adds a value of zero to the numerator of these formulas (Enders, 2006). This method also distorts covariances and inter-correlations between variables (Schafer and Graham, 2002).

After the problems that preserving the means instead of variances in a distribution is understood, survey methodologists have developed various imputation methods preserving distributional shape (Schafer and Graham, 2002). *Hot-deck Imputation* is one of these methods, where a respondent's value (donor) is selected at random, with this value being assigned to the non-respondent (Durrant, 2005). Basic idea is that, in order to impute a variable, a set of categorical variables are found that are associated with the one to be imputed. After a contingency table is formed based on the associated variables, and for the missing cells of the variable to be imputed, same values of other cases are used (Allison, 2002). Different strategies for selecting donor

respondents are also recommended as well as the stratification and sampling of donors (GSS, 1996). An advantage of the method is that actually occurring values are used for imputation. However, the method assumes that all units have an equal probability of response, and that units with missing data have similar characteristics to those with completed data. Selecting the donor variable in a proper way is another problem created by the method. Furthermore, it may distort relationships between variables (GSS, 1996). *Cold-deck Imputation* is a very similar technique to Hot-deck; the only difference is donor values are taken from previously conducted surveys. *Nearest neighbour imputation* is another donor method where the donor is selected by minimising a specified distance (Kalton and Kasprzyk, 1986). The observed unit with the smallest distance to the nonrespondent unit is identified and its value is substituted for the missing item according to the variable of concern. The same advantages and disadvantages hold for these methods as the hot-deck imputation.

Another broad class of methods for imputing missing data is *regression imputation* (Kalton and Kasprzyk, 1986). Predictive regression imputation, also called deterministic regression or conditional mean imputation, replaces missing variables by predicted values from a regression of the missing variables on variables observed for that unit. It involves the use of one or more auxiliary variables. A regression model is fitted that relates the dependent value to the auxiliary variables. The predicted values are used for imputation of the missing values. A potential disadvantage of predictive regression imputation is that it distorts the shape of the distribution and the correlation between variables, which are not used in the regression model (Durrant, 2005).

Under *random regression imputation*, also referred to as imputing from a conditional distribution, each missing value is replaced not by a regression prediction but a random draw from the conditional or predictive distribution. A residual term, thereby, is added to the predictive value to allow randomisation and to reflect uncertainty. A random regression model maintains the distribution of the variables

and allows for the estimation of distributional quantities (Kalton and Kasprzyk, 1986). An advantage of regression imputation is that it can make use of many categorical and numerical variables. A potential disadvantage of such a parametric model, on the other hand, is that the method may be sensitive to model misspecification of the regression model. If the regression model is not a good fit the predictive power of the model might be poor (Little and Rubin, 1986).

Many publications reviewed and presented comparisons of these classical methods. Pigott (2001) reviewed these widely used methods and argued that only complete case analysis provides valid estimates under the least number of conditions, and is applicable to a wider range of situations than available case analysis. He didn't recommend mean imputation under any circumstances. As another instance, Allison (2001) argued that listwise deletion is the least problematic, while all other conventional methods introduce bias into the standard error estimates. Conversely, Tanguma (2000) compared four methods; listwise deletion, pairwise deletion, mean imputation and regression imputation and said that it is best not to use the defaults on some of the statistical packages.

Bailar and Bailar (1978) compared the first two moments of estimated row means when missing values are a) ignored, or b) imputed by hot deck procedures; and found that under usual and/or reasonable assumptions both methods are unbiased. Durrant (2005) asserted that for a hot-deck method to work well a reasonably large sample size may be required, and the method is especially adequate when dealing with categorical data. Cox and Folsom (1978) performed simulations on different missing data techniques and reported that hot-deck imputations performed better than listwise deletion.

Cohen (2003) considered single imputation methods that are intended for use with standard variance formulas. He proposed that imputed values be more dispersed than

the observed values, thus compensate for the underestimation of variance by the usual formulas. Further, Cohen (1996), proposed that items be imputed from distributions more diffuse than those of the real data, thereby compensating for the underestimation of variance by the usual formulae. The approach is not intended for all statistical applications, only those based on the first two moments of means.

Yansaneh et al. (1998) asserted that the selection of an imputation method requires overall understanding of the nature of missing data in terms of extent and patterns. They proposed the following guidelines when deciding on an imputation strategy for large complex datasets:

a. Covariates cannot be missing when the imputation variable is missing, therefore some highly correlated variables that have missing values may not be appropriate as covariates,
b. When the variable to be imputed is continuous, regression imputation is used, if a highly predictive model – that is a high R-squared value – is obtained,
c. When the variable to be imputed is categorical, and weakly correlated with its covariates, there are two options: if the nonresponse rate is high, then Hot Deck Imputation is used, if nonresponse rate is low mean or modal imputation may be used.

Many authors (such as Allison, 2001 and Graham et al., 2003) discussed that all these classical methods introduce substantial bias and yield standard error estimates that are generally lower. In addition, they do not regard the imputation process involves uncertainty about the missing values (Allison, 2001 and Rubin, 1988).

2.3.2. Multiple Imputation

Rubin (1987) first proposed multiple imputation methodology to deal with missing data. In multiple imputation each missing value is replaced by a list of $m>1$ values. Substituting the jth element of each list for the corresponding missing value produces m plausible alternative versions of complete data (Schafer and Graham, 2002). Each of the data sets is analyzed in the same way by a complete case method. The results are then combined using techniques suggested by Rubin (1987) to give parameter estimates and standard errors that take into account the uncertainty due to missing values. Schafer (1999) stated that unless there are unusually high rates of missing data, the optimum is to use five to ten imputations. In many practical applications, the additional time and effort required to handle $m=20$ versions than $m=10$ has often little consequence (Schafer and Graham, 2002).

Multiple imputation was developed especially in the context of large survey studies which are to be used by a potentially large number of researchers for a number of different analyses (Sinharay et al., 2001). Thus once the multiple imputations are created, all users may analyze the resulting complete data sets using standard statistical software. Simulation studies have shown this method to be flexible and yield good standard errors, which are smaller than those obtained by other methods (Wayman, 2003, Rässler, 2004, Schafer and Graham, 2002).

Providing the imputations is often seen as the responsibility of the data provider (Schunk, 2006). This is firstly because imputation is a resource-consuming process that is not at the disposal of many users. Secondly, because some pieces of information are very important in imputation process, such as interviewer characteristics are not available to the public (Schunk, 2006).

Three assumptions are required for multiple imputation, as mentioned by Sinharay et al. (2001): (a) a model for the data values, (b) a prior distribution for the parameters of the data model, and (c) the nonresponse mechanism.

The first and the most crucial step in performing multiple imputations is to assume a probability model that relates the complete data Y, which is the combination of observed values and missing values in a dataset, to a set of parameters. Using this probability model and the prior distribution on the parameters, a predictive distribution for the missing values conditional on the observed values is obtained, and the imputations are generated from this predictive distribution (Sinharay et al., 2001). This assumed model should incorporate all the knowledge about the process of generating data and should be rich enough to preserve associations and relationships among variables that are of importance to the subsequent analysis (Schafer, 1999). For the continuous variables the most appropriate model is the multivariate normal assumption. Other models like log-linear can be used for other types of variables, however, the normal distribution still works well (Schafer and Graham, 2002).

According to Little and Schenker (1995), “multiple imputation can be motivated most easily from Bayesian perspective.” A Bayesian approach to a problem starts with the formulation of a prior distribution on the parameters to carry out the analyses, which is meant to capture the knowledge about the situation before seeing the data. After, the Bayes' Rule to obtain a posterior distribution for these unknowns is applied, which takes into account both the prior and the data (www.bayesian.org). From this posterior distribution predictive distribution of the missing values given the observed values is obtained.

As the third assumption, model-based multiple imputation assumes that the missing data is MAR, which was explained in the previous section. Schafer and Graham (2002) discussed that models which are not MAR are difficult to handle from a computational standpoint.

Rubin (1978) clarified the advantages of multiple imputation such that; "(1) imputing one value for a missing datum cannot be correct in general, because we don't know what value to impute with certainty; and (2) in order to insert sensible values for missing data we must rely on some model relating unobserved values to observed values" (Rubin, 1978). Multiple imputation accounts for missing data by restoring not only the natural variability in the missing data, but also by incorporating the uncertainty caused by estimating missing data (Rubin, 1988). Maintaining the original variability of the missing data is done by creating imputed values which are based on variables correlated with the missing data and causes of missingness. Uncertainty is accounted for by creating different versions of the missing data and observing the variability between data sets. Wayman (2003) mentioned another advantage of multiple imputation. It is very user friendly and familiar to many researchers, and it works in conjunction with standard complete-data methods and software. MI can also be used to fill in missing values in a multivariate missing data setting, and is suitable for numeric and categorical variables. It is currently probably the most practical and general approach, in particular for social scientists carrying out a large number of different analyses and missing values in several variables.

In practice, different ways exist on how to implement multiple imputations. *Markov Chain Monte Carlo* (MCMC), and especially data augmentation algorithms, defined in a Bayesian framework can be used for generating the missing data simulations (Wayman, 2003). In MCMC, a sequence of dependent random variates is generated whose distribution converges to the desired target. The algorithm provides imputed values from the conditional distribution of missing values given the observed values, where the distribution is integrated over any unknown parameters in the model with

respect to the posterior distribution of the parameters given the data. The multiple imputation by chained equations (MICE) method, also referred to as regression switching (Wayman, 2003). It enables the implementation of multiple imputation for non-monotone missing data patterns based on a sequence of regression models. Raghunathan et al. (2001) developed a sequential regression approach to multiple imputation. The idea is to regard a multivariate missing data problem as a series of univariate missing data problems. This latter method is used in the analysis of this dissertation, and detailed information on it is given in Methodology Chapter below.

In line with the different ways to implement multiple imputation, different software has developed. All the available procedures assume MAR. As the most frequently used statistical programme SPSS has a Missing Value Analysis (MVA) procedure, which calculates some basic statistics of variables subject to nonresponse and handles missing data based on some deletion and single imputation methods including regression imputation. STATA package includes options for various forms of hot-deck imputation techniques and options for MICE method. SAS offers various methods for missing data including regression and MCMC. IVEware is implemented in SAS and performs single and multiple imputations using the sequential regression imputation method (Raghunathan et al, 2001). NORM is a stand alone programme performs multiple imputations (Schafer and Graham, 2002). AMOS is developed to build attitudinal and behavioural models that reflect complex relationships in the dataset; and the software can impute numerical values for ordered-categorical and censored data based on Bayesian estimation, and estimate posterior predictive distributions to determine probable values for missing or partially missing data in a latent variable model as well (http://amosdevelopment.com/).

While multiple imputation appears the most promising of current missing data methods, Rubin (1996) critically reflects on the use of multiple imputation over the past 20 years. Some criticisms of the method center on the amount of computing and analysis time. A more critical assessment comes from Fay (1991, 1992) and is also

addressed by Meng (1994). Fay (1991) focuses on the use of multiple imputation in large, public-use data sets where the person imputing the data set is separate from the analyst. He proposed "bracketed response questions" for especially financial information questions, in order to reduce the rate of completely missing data, in his counter examples to multiple imputation.

2.4. Previous Applications of Multiple Imputation to Large Datasets

Missing values occur in many real world data sets, and some publications report on application of multiple imputation algorithms for missing values treatment. The following studies have some references to implementation of multiple imputation in the respective fields.

In 1992, a group of statisticians attempted to impute both item and unit nonresponses in the National Health and Nutrition Examination Survey (NHANES) of USA. A data file consisting of 27 key variables for 12,392 sampled adults was multiply imputed using techniques of iterative Bayesian simulation via Markov chains technique. This project represents the first successful implementation of proper multiple imputation methodology in a large multivariate survey and gave important insight about the future implementations (Khare et l., 1993).

The same dataset was studied also by Ezzati-Rice et al. (1993). The researchers compared three imputation methods and assessed potential imputation strategies for the NHANES III-Phase 1 data. They found that for the subset of data evaluated, values generated from two separate single imputation methods exhibited nearly identical distributions. In addition, the single and multiple imputation methods exhibited similar point estimates. Also, both methods preserved the marginal distribution of the variables and the relationship between them.

Similarly, in 1996, a model-based multiple imputation method was implemented in NHANES III for selected measurements, and the imputed dataset with 5 imputations is released separately from the original dataset (Schafer, 1996).

Statistical Commission and Economic Commission for Europe (1999) presented a model-based item imputation methodology intended as an alternative to the hot deck substitution algorithms for demographic surveys and censuses. The imputations produced with the methodology indirectly enjoy the properties of consistency and efficiency, in the sense that they are closely associated with consistent and efficient estimators.

In 2002, Taylor et al. (2002) applied multiple imputation to Flint Men's Study of African-American men about the prostate cancer and urologic symptoms. They found that multiple imputation corrected for nonresponse bias associated with observed data on age, and for other variables results from observed data and multiply imputed data were similar (Taylor et al., 2002).

Schenker et al. (2006) performed multiple imputation to handle missing data on family income and personal earnings in the National Health Interview Survey (NHIS). They described the approach used and concluded that imputation for biases that occur in estimates based on the data without imputation and imputation results in gains in efficiency as well (Schenker, 2006).

In 2006, Schunk (2007) applied MCMC multiple imputation procedure to a socio-economic survey of German Households, the SAVE survey. He also compared this procedure with hot-deck and regression methods.

In addition to the above applications, there are few applications of imputation techniques to DHS surveys. First one is concerned with the imputation of marital status as a determinant of living arrangements of older persons, in a study investigating patterns and trends in the living arrangements of older persons for more than 130 countries (UN, 2005). It has been found that the inferences about the direction and statistical significance of effects are identical for selected variables regardless of whether the observed or imputed marital status is employed. In another study on the causal link between democracy and greater primary education provision, an imputation model is used in order to make the most efficient use of information available about primary school attendance (Stasavage, 2005)) The study also attempted to avoid potential biases introduced by missing data by multiple imputation technique applied.

2.5. DHS Editing and Imputation Procedure

Good quality demographic and health data is one of the main goals of Demographic and Health Surveys (DHS) Program (Croft, 1991). Although the complex design and questionnaire structure of the DHS' aim to minimize problems regarding missing data, some major variables are exposed to missing data and inconsistency problems. Special codes are used in DHS data for certain response types, including inconsistencies and missing values. The general coding scheme of DHS is given in Table 2.1 below. The codes are applied to four digit, three digit, two digit and one digit variables, respectively (Macro International, 2004).

Table 2.1 Special Codes Used by the DHS

Codes	Description
BLANK	Variable is not applicable for this respondent either because the question was not asked in a particular country or because the question was not asked of this respondent due to the flow or skip pattern of the questionnaire.
9999, 999, 99, 9	The question should have been answered by the respondent, but no information was available (missing data).
9998, 998, 98, 8	The respondent replied "Don't know" to this question.
9997, 997, 97, 7	The answer to this question was inconsistent with other responses in the questionnaire and it was thought that this response was probably in error. The response was changed to this code to avoid further problems due to inconsistency of information.

Source: Macro International, 2004.

There are some important variables where missing data is not accepted by DHS (Rutstein and Rojas, 2003). These variables are;

- Geographical variables such as urban/rural,
- Level of education for women and men,
- Current use of contraception for women,
- Current marital status of women, and,
- Some of the variables related to the women's birth history.

Currently, some editing and imputation techniques are used to handle these problems. The DHS approach to editing of data foresees three distinct phases, which are editing during data entry, secondary data editing and imputation phases. Editing during data entry is restricted to controlling the structure of the data file, the skip patterns through the questionnaire, the range of valid values for each variable and the consistency of certain variables. During secondary editing stage complex checks are introduced to verify the internal consistency of information throughout the questionnaire. In the imputation stage, a new data file is produced in which partial or incomplete dates are imputed from the known related information, which is called recode data file.

Key dates are imputed when it is not provided by the respondent or in some cases if they are inconsistent (e.g. less than 7 months between two births). These key dates are; date of birth of the respondent, date of first union or marriage, date of birth of each child of the respondent, date of conception of current pregnancy, date of sterilization of respondent or partner and date of start of use of current method. Only imputed dates are available in the recode data file. However a date flag is included to show what format the information was prior to imputation. The codes for this date flag are as follows (Macro International, 2004):

1. Both month and year of the event were given and so no imputation was necessary.
2. The year of the event was not given, but the month of the event and the age of the respondent or child or, in the case of the date of first union, the respondent's age at first union were given. In most cases this information uniquely identifies the exact date of the event. In a few cases the year of the event was imputed from a choice of two possible years.
3. The year of the event, but not the month, and the age of the respondent or child or, in the case of the date of first union, the respondent's age at first union were given and only the month of the event was imputed.
4. The year of birth, but not the month, and the age of the respondent or child were given. However, in surveys where it is believed the year of birth is calculated from the age, the year of birth is ignored when the year of birth plus the age add up to the year of interview.
5. The year of the event was given but the month of the event was not given, and neither was the age. The month of the event was imputed.
6. Neither the month nor the year of the event was given, but age was given and the year and month of the event were imputed from the age.
7. Only the month of the event was given, without the year or age. The year of the event was imputed from other available information.
8. No information was given concerning the date of the event; however month and year of the event were imputed from other information.

The method used in the DHS program for imputation of dates is based on the construction of logical ranges for each date, which are refined in three stages. As the first stage, an *unconstrained range* is developed from the available information, which consists the earliest and the latest possible dates at which the event can have occurred. If only a year is available, the unconstrained range spans 12 months. If no year is given the unconstrained range covers the full range of possible dates, i.e. 50 years before interview until 5 years before interview for the date of birth of the respondent.

As the second stage, ranges are adjusted in the light of *isolated constraints*, which are items of data concerned to a particular event, but with no relation to any other event. The isolated constraints in DHS include age of the respondent, age of each child, duration of current pregnancy and age at death of children.

At the third stage of imputation, ranges are adjusted to satisfy *neighboring constraints*, which are restrictions placed upon the range of acceptable dates by earlier and later events in the respondent's life. Gestation length of a pregnancy and durations of amenorrhea, abstinence and breastfeeding after the birth of a child are some examples for neighbouring constraints.

At the end of this process, the difference between upper and lower bounds are obtained for each event. If this range is negative, then the date is inconsistent. If range is zero, the date is consistent with other related information, and the constraints were sufficient to restrict it to one month. Finally, if the date is positive, the date is consistent with other related information but incomplete, since the constraints were insufficient to restrict it to one month. In such a case, a random imputation method is used to assign the imputed data within the final logical range for each event (Croft, 1991).

According to Croft (1991), however, there are several problem areas emerge at the imputation process of different waves[2] of DHS surveys. A few examples are reported by Croft (1991) concerning the first two waves of DHS, which are listed below:

In first DHS I, there was a cut-off for the inclusion of children for the health related questions, which was also used in imputation of dates of birth of children according to some procedure. However, for children without a year of birth, most interviewers tended to exclude them from health related section. Hence, some surveys are affected very seriously in terms of bias in a number of indicators. This constraint has been dropped in DHS II due to these biases.

Ancillary data such as durations of breastfeeding, amenorrhea and abstinence after the birth and the duration of contraceptive use between two births were used to constrain the bounds of the dates of births in DSH I. However these data were particularly prone to heaping, and produced inconsistent data.

Two obvious biases associated with the DHS imputation procedure for the dates were birth intervals and pre-marital births. Considering the birth intervals for individuals the imputation process produced short birth intervals and under or over-estimated.

Some changes were made to correct the problems in previous waves; however improvement areas exist in editing and imputation procedure of DHS (Croft, 1991). Croft (1991) concludes that "the need for data editing and imputation techniques serves to indicate that there is still a long way to go.".

[2] First wave of DHS is called as DHS I which covers years 1984-1989, second wave is called DHS II and covers years 1988-1993, third wave is called DHS III and covers years 1992-1997, and currently implemented DHS' are called MEASURE DHS which covers years 2003-2008, (http://www.measuredhs.com/aboutdhs/history.cfm).

2.6. Studies Related to Data Imputation in Turkey

The problem of missing data has recently attracted much interest with the growing concerns about the issue worldwide recently. However there is an obvious lack of both theoretical and practical studies in Turkey. The institutions, that are conducting large scale surveys as well as censuses, do not have established imputation methods, although some ad-hoc imputation techniques are applied. As well as that, there are a very limited number of articles in Turkey.

In his Ph.D. dissertation, Türkyılmaz (2003) used small area estimation techniques in order to obtain provincial estimates of selected demographic and health indicators, by using 1990 Census and TDHS-2003 datasets. He multiply imputed the estimates of districts where the census variables assumed as the fully observed variables and survey variables assumed to be variables with missing in an aggregated district level dataset. In this way he calculated the standard errors, that couldn't be calculated by the small area estimations technique. He concluded that, while the estimations for high prevalence indicators produced reliable estimates of multiple imputations and composite estimates, the reverse holds for the low prevalence indicators (Türkyılmaz, 2003).

A simulation study is conducted in which different imputation methodologies are compared by using datasets derived by a software program with different sample sizes. It is reported that, listwise, pairwise, mean and regression imputation methods have consistency problems for sample sizes less than 200 (Bal and Özdamar, 2004).

Oğuzlar (2001) first reviewed the existing literature on imputation methods. She also used a dataset obtained from the World Bank and excluded variables that have

missing values over 60% rates. She applied listwise, pairwise and regression imputation techniques on SPSS MAV module and found that the resultant dataset is consistent (Oğuzlar, 2001).

3. METHODOLOGY

3.1. Data Source Used in the Study

In this section, TDHS-2003 as the main data source used is introduced in terms of a brief history, sampling design and the quality of the data in relation with the missing data. In addition, overall existence of missing data in TDHS-2003 is investigated prior to the discussion on anthropometric measurements as the study variables.

3.1.1. The 2003 Turkey Demographic and Health Survey

The main data source used in the study is the 2003 Turkey Demographic and Health Survey (TDHS-2003), which was implemented by the Hacettepe University Institution of Population Studies (HUIPS). TDHS-2003 is the third and latest survey implemented by employing the DHS methodology.

TDHS-2003 provides data on socioeconomic characteristics of households and women, fertility, mortality, marriage patterns, family planning, maternal and child health, nutritional status of women and children, and reproductive health. The data collected is of vital importance for Turkey, since they are the only source from which many nationwide population and health indicators are generated. The data collected in TDHS' are intended for the utilization of policy makers to evaluate and improve family planning and health programs in Turkey. Furthermore, sustaining flow of information for the related governmental and other organizations in Turkey as well as abroad on the Turkish population structure in the absence of vital registration system is another key contribution of the TDHS' (HUIPS, 2004).

Specific objectives of TDHS-2003 are as the following (HUIPS, 2004);

- To collect nationally representative data that allows the calculation of demographic rates, particularly fertility and childhood mortality rates;
- To obtain information on determining direct and indirect factors of levels and trends in fertility and childhood mortality;
- To measure the level of contraceptive usage and usage by method type, region, and urban-rural residence;
- To collect data on mother and child health, including immunizations, antenatal care, assistance at delivery and breastfeeding;
- To measure the nutritional status of children under five and of their mothers; and
- To collect data on elderly welfare, knowledge of sexually transmitted diseases and AIDS, as well as usage of iodide salt.

Two main types of questionnaires were used in the TDHS-2003, which are the Household Questionnaire and the Individual Questionnaire. The Household Questionnaire was used to enumerate all members of and visitors to the selected household and to collect information on the socio-economic level of the households. Basic information, including age, sex, educational attainment, marital status, working status and relationship to the household head, on each person listed as a household member or a visitor is collected in the first part of the Questionnaire. The main objective of this first part was to identify eligible women, in terms of ever-married and in 15-49 age group, as well as to obtain the general socio-economic and socio-demographic profile of the Turkish households. The second part of the household questionnaire was designed to collect data on welfare of elder population, if any, in the households. In the third part, questions on the dwelling unit and the ownership of a variety of consumer goods are included. In this part, Istanbul Metropolitan Household Module was also included to collect data on tenure, availability of electricity, piped-water and natural gas in households located in the urban residences. In the final part of household questionnaire, questions on the use and storage of the salt used for cooking are included. These salt-related questions are asked in the half

of the sampled clusters, and salt iodization tests were applied in the interviewed households in these clusters (HUIPS, 2004).

In the Individual Questionnaire, the areas of data collection were; background characteristics, birth history, marriage, knowledge and use of contraceptive methods, other information related to contraception, abortions and causes, maternal health care and breastfeeding, immunization and acute respiratory infections, fertility preferences, husband's background characteristics, women's work and status, knowledge of sexually transmitted diseases and AIDS, maternal and child anthropometry. In addition, a calendar module in the Individual Questionnaire was used to record fertility, contraceptive use and marriage events on a monthly basis for six and a half years beginning from January 1998 up to survey month.

A weighted, multistage, stratified cluster sampling approach was used in the selection of the TDHS-2003 sample (Türkyılmaz et al., 2004). The major objective of the TDHS-2003 as well as all DHS' sample design is to ensure that the survey would provide estimates with acceptable precision for the domains for most of the important characteristics, such as fertility, infant and child mortality, and contraceptive prevalence, as well as for the health indicators. The sample design and sample size of TDHS-2003 make it possible to perform analyses for Turkey as a whole, for urban and rural areas and for the five demographic regions (West, South, Central, East and North). In addition to these regions, TDHS-2003 sample size allows analyses on some survey topics for the 12 NUTS1 regions. Among these, Istanbul and Southeastern Anatolian Project Regions (GAP) were over-sampled for specific analysis (Türkyılmaz et al., 2004).

Fieldwork of the TDHS-2003 began by December 2003 and was completed by May 2004. The results of sample implementation indicate that, a total of 11.659 households were located and visited, of which 10,836 households were successfully

interviewed. Overall, the household response rate was calculated as 93 percent. The household response rate was higher in rural areas than in urban areas, and highest in East, North and South regions. HUIPS (2004) reported that, the response rates in Istanbul were the lowest with 84 percent where it is more than 98 percent in Northeast Anatolia. In addition, the overall response rate for women was calculated as 89 percent, ranging from 83 percent in the Central region to 93 percent in the East region (Türkyılmaz et al., 2004).

The "Quality of the Data" section of the TDHS-2003 Final Report deals with inconsistencies and missing information on some key variables (Koç, 2004). First of all, *heaping* was observed in the reporting of ages ending with 0 and 5, especially in the older ages. According to Koç (2004), the results does not show any evidence that interviewers 'aged' children out of the eligible range for the collection of height and weight and health data. Because the proportion of children reported to be five years of age at the time of the survey is almost equal to the proportions age four and six. It appears to have been little shifting of older women past age 49, which is the upper limit of eligibility of individual interviews. In addition, response rates were found to be lower for oldest and youngest age groups indicating that interviewers may have been somewhat less diligent in pursuing interviews with women at the two extremes of eligible age range. Another quality indicator which is also focal point of this thesis is the extent to which information is missing on key variables, including birth date and anthropometric measurements, which are most prone to missing data problems (Koç, 2004).

In Table 3.1 below, information on the completeness of reporting in connection with a set of important variables for the last three TDHS' is provided. From the table it can be seen that, height and weight measurements were missing for approximately 10 percent of the children under age 5 in TDHS-1993 and 18 percent in TDHS-1998. Missing information in height and weight measurements declined slightly to 8 percent in TDHS-2003, which is a favourable result.

Table 3.1 Completeness of reporting (weighted) for selected variables in TDHS-1993, TDHS-1998 and TDHS-2003 data sets

Subject	TDHS-1993		TDHS-1998		TDHS- 2003	
	% missing information	Number of cases	% missing information	Number of cases	% missing information	Number of cases
Birth date*						
Month only	2,05	12639	9,4	10368	3,7	12646
Month and year	0,24	12639	0,8	10368	1,2	12646
Anthropometry **						
Weight missing	9,36	3532	12,8	3299	5,5	3998
Height missing	7,61	3532	16,9	3299	7,3	3998
Weight and height missing	9,45	3532	17,6	3299	7,6	3998

* For births in the 15 years preceding the survey.
**For the living children 0-59 months.

3.1.2. Missing Data in TDHS-2003

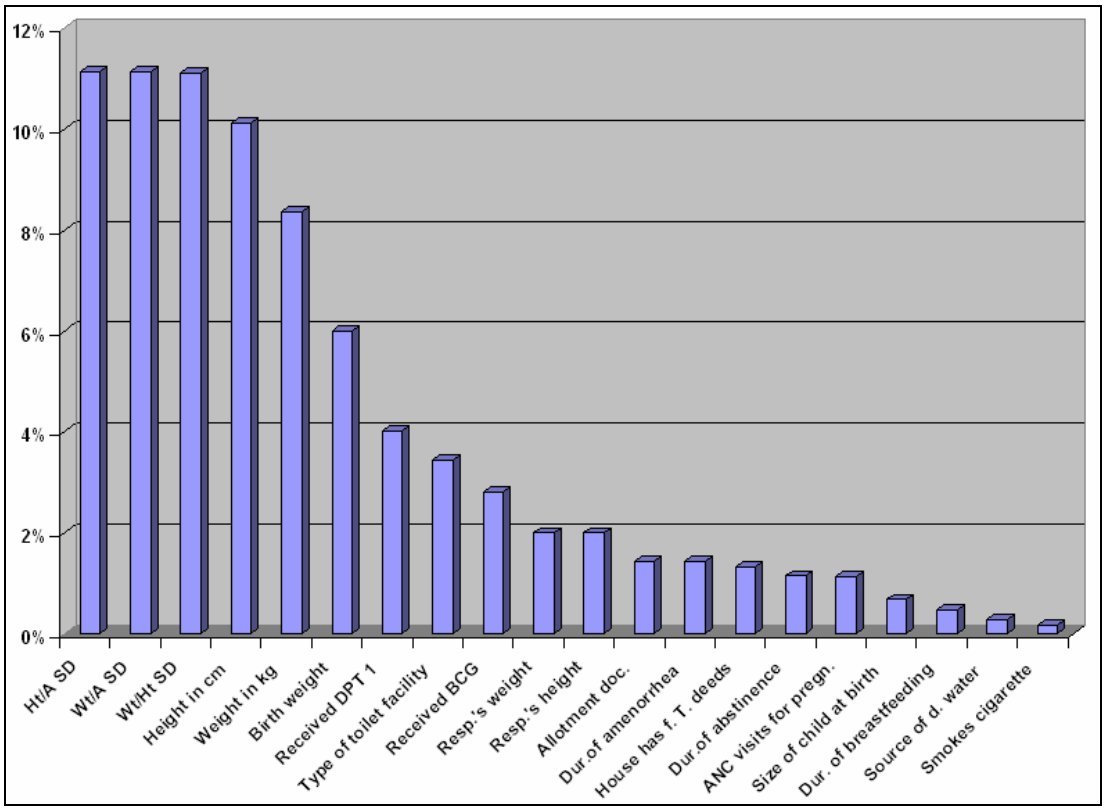
Percentage of missing values in some selected variables of the TDHS-2003 is given in Figure 3.1 below. In order to facilitate interpretation, it should be noted here that, according to the literature, missing data rates of less than 1% are generally considered trivial and 1-5% manageable; however, 5-15% require sophisticated methods to handle, and more than 15% may severely affect any kind of interpretation (Acuna and Rodriguez, 2004). When the TDHS-2003 data is examined, it is seen that for the majority of the variables, missing data rates don't exceed 5 percent. However, missing values generally occur in measurement related questions.

In addition to that, there are serious discrepancies among different representation groups (regions or urban/rural), which probably cause problems for specific analyses at respective levels. As a remarkable instance, while the overall missing rate for the "month of birth" variable is 22%, it increases to 36% for rural residences and 39% for East region[3]. Similarly for anthropometric measurements for children under age

[3] Analysis is made from the Raw Data, which is in the form of collected from the field and not imputed by the DHS (variable q105m).

five, 8% overall missingness rate in weight increases up to 11% for North region; and 10% overall missingness rate increases to 11% for North and West regions.

Figure 3.1 Percentage of Missing Values in Selected Variables in TDHS-2003 (unweighted)



As mentioned earlier under *DHS Editing and Imputation Procedure* section, there are different kinds of codes used in DHS data files, including BLANKS (skipped questions), inconsistencies, don't knows and missing values. Schafer and Graham (2002) opened a debate on which of these codes should be evaluated under "missing values to be imputed". They stated that most of the researchers would not consider the skipped items to be missing, however, in the presence of response error; it might be a mistake to presume that all skipped items are zero, because some answers to the initial question may be incorrect. Another issue they have mentioned is on longitudinal studies in which unfortunate (like deaths) events preclude measurement. They assert that if deaths are unrelated to the subject of interest, then parameters may be estimated for an ideal scenario in which no one dies during the study. Thus in

some contexts where MAR assumption is not prevented, accepting the BLANKS as missing values is reasonable (Schafer and Graham, 2002). These contexts definitely do not cover the cases for whom the question is not applicable; as an instance, number of births for men. On the other hand, cases for whom the question is applicable, but that are not able to response it for some unfortunate events can be covered. The reason for that is the social events can be explained by some other dependent conditions for those latterly mentioned; such as the number of births for a women living in specific conditions. The decision of involving these kinds of cases depends on the purpose of the study.

In the Individual Data of TDHS-2003, there are 4533 children of whom 161 are dead. Naturally, height and weight measurements of these children are not taken as well as some other data, and coded BLANK in the data file. In Figure 3.1 above, missing data is presented in the context of all kinds of nonresponses including BLANKS, missing values and don't knows. 161 dead children are included in this figure for just illustration purposes, in terms of showing the magnitude of missing data according to the above mentioned approach. The first three variables which are standard deviations for height for age, weight for age and weight for height; reflect the inconsistencies in weight, height and age variables. As seen from Figure 3.1, considerable portion of weight and height data is flagged, because of inconsistencies. Although all kinds of response problems are covered in Figure 3.1, BLANKS are out of the scope of this study. In this context, missing values, inconsistencies and don't knows, in the form they are defined under *DHS Editing and Imputation Procedure* section, are regarded as missing values throughout this study.

DHS has a well-grounded imputation scheme for the key events, as explained earlier in this study. These key dates in a human life is difficult to remember, thus dates are naturally very prone to be missing. In line with this logic, dates are the variables, which include missing values at highest rates, when the TDHS Raw Data is examined. Indeed, DHS imputation scheme allows a very narrow range for a random

imputation, after all logical constraints. In addition if the date variables were supposed to be the dependent variables in an imputation model, there were not many covariates, which could be used in predicting the missing information, except for other date information. However, the major aim of this study is to test multiple imputation technique on some variables conditional on other observed variables. On the basis of this rationale, missing data on dates are considered out of scope of this thesis, and imputed data files are used.

3.2. Anthropometric Measurements

One of the most important contributions of TDHS' data is the anthropometric measurements which provide a useful assessment of the nutritional and health status of a population. When missing values are investigated in order to test the imputation methodology in TDHS-2003; it is seen that missing data specifically concentrates on anthropometric measurements of children under age five. In line with this outcome, anthropometric variables, namely weight and height are chosen as the core study variables to be imputed in this study, based on several reasons. Firstly, anthropometric indexes play an important role in monitoring and evaluation of the development progress of countries. *Underweight*, as one of these indexes was chosen as the key indicator for the first Millennium Development Goal (MDG)[4] for "eradication of extreme poverty and hunger". Another reason is the appropriateness of type of anthropometric data (i.e. continuous) for the development of the imputation model is an advantage to test the methodology proposed.

Anthropometric data has a meaning beyond predictive purposes. Cogill (2003) asserts that well-chosen and reported indicators would enhance program management as well as providing valuable insights. Moreover, he classified the topics on which

[4] Millennium Developed Goals are developed by the United Nations in 2000, which are eight goals to be achieved by 2015 that respond to the world's main development challenges (http://www.undp.org/mdg/basics.shtml).

anthropometric data can serve. These topics include; identification of individuals or populations at risk, selection of individuals or populations for intervention, evaluation of the effects of changing nutritional, health or socio-economic influences; excluding individuals from high risk treatments, from employment (a job requiring physical strength) or from certain benefits; achieving normative standards and research purposes with other aims than nutrition and health. In general, anthropometry is very relevant to program management because of three reasons, (i) to identify target groups, (ii) in monitoring progress, and (iii) in assessing overall program effects (http://www.unsystem.org/SCN/archives/npp07/ch10.htm).

The TDHS-2003 data include anthropometric measurements for each respondent's children under 5 years of age. Collection of anthropometric data requires additional steps and resources in survey studies, as well as TDHS'. In TDHS', interviewers deliver a specific training on anthropometric measures, before field work. Special equipment is provided for measurements and additional time is spent during data collection. In other words, anthropometric data collection is a costly work. However, missing values as well as inconsistency problems generally occur at high rates in anthropometric measurements due to some uncontrollable reasons. Making use of all relevant available information is a reasonable and useful method to deal with missing values in order to get better estimates.

There are four pillars used to make anthropometric assessment, which are sex, age, weight and height (Cogill, 2003). When these variables are used together, they can provide important information about the nutritional status of a person or a population. When two of these variables are used together, it is called an index. Three indexes are used in assessment of children's nutritional status; which are height for age, weight for height and weight for age. These indexes are compared with a reference population for interpretation. The reference standards most commonly used to standardize measurements were developed by the U.S. National Center for Health Statistics (NCHS), the Center for Disease Control and Prevention

(CDC) and World Health Organization (WHO), and called NCHS/CDC/WHO reference. This reference population was recommended by the World Health Organization (WHO) for international use (Cogill, 2003). The reference population chosen by NCHS was a statistically valid random population of well-nourished and healthy children. There are debates on the validity of the US based reference standards for populations from other ethnic backgrounds recently. Some researchers assert that until the age of 10 years, children from different backgrounds grow approximately the same rate (Cogill, 2003). DHS Surveys use NCHS/CDC/WHO international references standard, in comparisons.

References are used to standardize a child's measurement by comparing it with the median or average measure for children at the same age and sex. In describing the differences from the reference, a numeric value is standardized to enable children of different ages and sexes to be compared. Differences in measurements can be expressed in various ways; (i) standard deviation units (or Z-scores), (ii) percentage of the median, and (iii) percentiles.

Z-scores is defined as the difference between the value for an individual and the median value of the reference population for the same age or height, divided by the standard deviation of the reference population, which can be displayed as the following;

$$\text{Z - score} = \frac{\text{(observed value) - (median reference value)}}{\text{standard deviation of reference population}}$$

The percentage of the median is defined as the ratio of a measured value in the individual to the median value of the reference data for the same age or height for the specific sex, expressed as a percentage, which can be written as follows;

$$\text{Percent of median} = \frac{\text{observed value}}{\text{median value of reference population}} * 100$$

The percentile is the rank position of an individual on a given reference distribution, expressing what percentage of the group the individual equals or exceeds.

The distribution of Z-scores follows a Normal distribution. The most commonly used cut-off with Z-scores is "-2 standard deviation" (-2 SD); which means with a Z-score for underweight, stunting or wasting below -2 SD are considered moderately or severely malnourished. In the NCHS/WHO/CDC classification system, cut-offs which help analyze and present data are as the following;

Table 3.2 NCHS/WHO/CDC Cut-off Values in Malnutrition Classification

	Severe - 3.01 or below	Moderate -2.01 to - 3.00	Mild - 1.01 to - 2.00	-1.00 to +1.00	+1.01 to +2.00	Over nourished +2.01 or above	Total
Expected percent	0.1	2.2	13.6	68.2	13.6	2.3	100.0

Source: Macro International (2007).

The extent to which the percentage of children falling into these categories exceeds 2.3% (the expected percentage in a well nourished population) indicates the level of specific aspects of malnutrition in the population. The three indicators of malnutrition are commonly constructed from anthropometric measures, which are;

(1) an indicator of *stunting*, a condition reflecting chronic malnutrition and designated by a low height for age;

(2) an indicator of *wasting*, a condition reflecting acute or recent malnutrition and designated by a low weight for height; and

(3) an indicator of *underweight*, a condition reflecting chronic and/or acute malnutrition and designated by a low weight for age.

In any large population, there are natural variations in height and weight. These variations approximate a normal distribution with the following percentages found in each standard deviation category.

In DHS analyses, a child is classified as "stunted" if s/he is below minus two standard deviations (-2 SD) from the reference median in terms of height for age, as "wasted" if below -2 SD from the reference median in terms of weight for height, and "underweight" if of a low weight for age by the same criterion.

In the process of assigning Z-scores, necessary checks are made on their plausibility. Z-scores are assigned missing to children with incomplete date of birth, based on the reason that z-scores are very sensitive to changes in age. Children with height for age z-scores below -6 or above +6 SD, with weight for age z-scores below -6 or above +6 SD, or with weight for height z-scores below -4 or above +6 SD are flagged as having invalid data. In addition, invalid combinations of z-scores where height for age is less than -3.09 SD and weight for age is more than -3.09 SD, or where height for age is more than -3.09 SD and weight for age is less than -3.09 SD.

Different software exists to make comparisons with the reference standards, where the most popular ones are EpiInfo and ANTHRO (Cogill, 2003). The WHO ANTHRO Software is used by the DHS'; in the calculation of anthropometric indexes and their comparison with the NCHS/CDC/WHO reference population. The software was developed to assess child nutritional status, to follow a child's development and growth over time, or to conduct and analyse nutritional surveys. In 2005 an MS Windows-based version of ANTHRO (ANTHRO 2005) replaced the previous ANTHRO 1.02 DOS-based one, which can also apply recently developed WHO Child Growth Standards as well as NCHS/CDC/WHO reference standards (www.who.int/childgrowth/software/en). Both versions give exactly the same results, however the newer version has additional features like tabulation and graphical

representation. ANTHRO 2005 is used for the calculation of anthropometric indexes and their comparison with the NCHS/CDC/WHO reference population in this thesis.

3.2.1. Missing Data on Anthropometric Measurements

A Multinominal Logistic Regression analysis is conducted to find out if the missingness is related to other variables in the TDHS-2003 dataset. In the analysis, "Reason not measured" variable, which is replied by the interviewer according to the result of height and weight measurements, is used as the dependent variable to investigate if other characteristics of respondents are in relation with the reason of not measurement. The frequency distribution of "reason not measured" is given in table below.

Table 3.3 Frequency Table for the Variable "Reason not measured", TDHS-2003

	Frequency	Percent
Measured	3696	92.44
Not present	49	1.22
Refused	89	2.22
Other	148	3.71
Sub-Total	3982	99.59
Missing	16	0.41
Total	3998	100.00

According to the results of multinominal logistic regression, the regression model is significant in statistical terms, as seen in table below. In addition, the likelihood ratio tests table for each variable is given below. According to the results, age in months, education of the respondent in single years, sex of household head, smokes cigarettes and wealth index affect the reason that the anthropometric measurements are missing (sig.<0,05).

Table 3.4 Model Fitting Table

	Chi-Square	Sig.
Model	150.30	0.00

Table 3.5 Likelihood Ratio Tests for selected variables

Effect	Definition	Chi-Square	Sig.
MONTHS	Age in months	23.65	0.00
V133	Education	43.49	0.00
V101	Region	9.40	0.67
V102	Urban/rural	6.94	0.07
SEX	Sex of child	1.11	0.77
V151	Sex of hh head	3.66	0.30
V463A	Smokes cigarettes	13.31	0.00
V190	Wealth index	28.90	0.00
SIZECHIL	Size of child at birth	10.87	0.54
MOTTONG	Mother tongue	10.40	0.11

The missingness of anthropometric data is related to several characteristics; therefore the measured sample cannot be treated as a random subset of the original sample, which means that missingness mechanism of anthropometric variables are not Missing Completely at Random (MCAR), according to the terminology.

3.3. Estimation Techniques Used in the Study

The imputation method used in this study is multiple imputation. Multiple imputation is an attractive approach when the dataset is used by diversified users, because once the missing values have been imputed, standard software can be used to analyze the completed data. As mentioned earlier, multiple imputation method enables calculating the within and between variance components due to imputation. Theoretical underpinnings of multiple imputation is given firstly under this section. After, selected multiple imputation method is introduced. Sequential regression

multivariate imputation (SRMI) developed by Raghunathan et al. (2001) is used to create the multiple imputations in the study. This choice of method is based on several advantages of the procedure which are discussed below. According to this procedure, imputations are created by using a special module under IVEware software package. This software is introduced under the subsequent section of the study, namely *Software Packages Used*.

3.3.1. Multiple Imputation

The basic idea of multiple imputation is as follows; impute the missing values using an appropriate imputation model that incorporates random imputation, repeat this M times, carry out the analysis of interest, e.g. the estimation of a proportion, in each of the M resulting datasets and combine the estimates using Rubin's rules (Rubin, 1988). According to this procedure, multiple imputations need to fulfil certain conditions which is referred to as proper multiple imputation. Rubin (1988) defines proper multiple from a frequenist perspective without reference to any specific parametric model. Applying proper multiple imputation enables the use of the resulting M complete data sets for performing standard complete-data analysis, combining the results for a single overall inference. The nice feature is that the differences in the M results obtained from the M complete-data set can be used as a measure of uncertainty caused by missing data.

Suppose that M is the number of imputations and e_l are the estimates from the imputed data set $l = 1,2,...,M$. Also let v_l be the corresponding variances of the estimates which are the squares of the standard errors. Multiply imputed estimate is;

$$e_{MI} = \frac{1}{M}\sum_{l=1}^{M} e_l$$

To obtain the multiply imputed variance estimate, firstly two variance estimates are computed, which are within-imputation variance (mean variance) and between-imputation variance (Raghunathan et al., 2002). *Within-imputation variance* is the average of the complete data variance estimates, and calculated as the following;

$$\bar{v}_M = \frac{1}{M}\sum_{l=1}^{M} v_l$$

and the variance estimate of the complete-data point estimates, defined as the *between-imputation variance* is calculated as the following;

$$B_M = \frac{1}{M-1}\sum_{l=1}^{M} (el - eMl)^2$$

Combining both forms of the variance estimates including an adjustment term $(\frac{M+1}{M})$ for finite M, defines the overall variance estimate associated with e_{Ml} is;

$$v_{Ml} = \bar{v}_M + \frac{M+1}{M} B_M$$

The total variance of the estimate is made up of two components, including a component which preserves the natural variability and an additional component which estimates uncertainty caused by missing data. In addition, R is termed as the fraction of information about *e* that is missing due to nonresponse, where

$$R = \frac{M+1}{M}\frac{B_M}{v_{Ml}}$$

which is also known as “fraction of missing information” (Schenker et. al., 2006).

3.3.2. Sequential Regression Multivariate Imputation (SRMI) Technique

Raghunathan et al. (2001) introduced the sequential regression multivariate imputation (SRMI) procedure for relatively complex data structures, which is based on the Bayesian approach. The Bayesian approach specifies an explicit model for the variables with missing values, which is conditional on the fully observed variables and some unknown parameters; a prior distribution for the unknown parameters; and a model for the missing data mechanism, which does not need to be specified under an ignorable missing data mechanism. This explicit model generates a posterior predictive distribution of the missing values conditional on the observed values. The imputations are draws from the posterior predictive distribution of the missing values given the observed values.

The "imputes" are created through a sequence of multiple regressions in the multivariate multiple imputation procedure. Covariates include all other variables observed or imputed on that variable. The sequence of imputing missing values are continued in a cyclic manner, each time overwriting the previously drawn values in order to build interdependence among the imputed values and exploit the correlational structure among covariates.

The type of imputation in the model changes by the variable being imputed, which can be in the form of (1) continuous, (2) binary, (3) categorical, (4) counts, and (5) mixed (a continuous variable with a non-zero probability mass at zero). Following models are used for the conditional regressions in the model (Rahgunathan et al, 2001);

1. A normal linear regression model on a suitable scale if the dependent variable is continuous,
2. A logistic regression model if the dependent variable is binary,

3. A Polytomous or generalized logit regression model if the dependent variable is categorical,
4. A Poisson loglinear model if the dependent variable is a count variable, and
5. A two-stage model where non-zero status is imputed using a logistic regression model and conditional on non-zero status, a normal linear regression model is used to impute non-zero values, if the dependent variable is mixed.

The imputation procedure also considers some other features of survey data, which generally make the modelling process difficult. These consist of former question restrictions, and logical and consistency bounds for the missing values.

To illustrate the theoretical framework of the procedure, let X denote the predictor matrix with no missing values, for a sample of n observations. X consists of the continuous, binary, count or mixed variables with no missing values and appropriate dummy variables representing the categorical variables. Moreover, it may also consist of a column of ones to model an intercept parameter, offset variables and certain design variables. Similarly, let Y_1, Y_2, ..., Y_k denote k variables with missing values, ordered by the amount of missing values from least to most. Then the joint conditional density[5] of $Y_1, Y_2, ..., Y_k$ given X, can be written as;

$$f(Y_1, Y_2, ..., Y_k \mid X, \theta_1, \theta_2, ..., \theta_k) = f_1(Y_1 \mid X, \theta_1)\, f_2(Y_2 \mid X, \theta_2) \ldots f_k(Y_k \mid X, \theta_k)$$

where f_j, j=1,2, ...,k are the conditional density functions. Each conditional density is modelled through an appropriate regression model with unknown parameters, θ_j, and values are drawn from the corresponding predictive distribution of the missing values given the observed values.

Each imputation consists of c rounds. In round 1, the variable with least missing values, namely Y_1 is regressed on X and the missing values are imputed under the

[5] The conditional density function describes the probability over a random variable given the value of another random variable.

regression model. X is then updated by appending Y_1 appropriately before moving to variable Y_2, with the least missing values and the same process is repeated. The imputation process is continued until all the variables have been imputed. the whole rounds are repeated *pM* times, where p refers to iterations and M refers to multiples in the imputation process. *p* value makes the imputed values be uncorrelated and *M* value defines the number of multiples.

Türkyılmaz (2003) illustrated these steps in a simple example consisting variables; x_1, x_2, y_1 and y_2; where the first two have no missing values, and last two have missing values from least to most. Round 1 starts with the following equation:

$$\hat{y}_1 = \beta_0^{(1)} + \beta_1^{(1)} x_1 + \beta_2^{(1)} x_2 + \hat{\varepsilon}$$

X, which is the data without missing values, is updated by appending y_1 accordingly. Then the process is repeated for y_2, the variable with the next fewest missing values, by adding the imputed variable y_1 to the model:

$$\hat{y}_2 = \beta_0^{(2)} + \beta_1^{(2)} x_1 + \beta_2^{(2)} x_2 + \beta_3^{(2)} y_1 + \hat{\varepsilon}$$

This procedure would continue for each of the missing values in the data set. Once this procedure is finished, each missing value has an imputed value substituted for it, resulting in a fully-completed data set, number 1.

From rounds 2 through *c*, the process is repeated, by modifying the predictor set to include all Y variables except the one used as the dependent variable. Thus, the two equations in round 2 can be illustrated as follows:

$$\hat{y}_1 = \beta_0^{(1)} + \beta_1^{(1)} x_1 + \beta_2^{(1)} x_2 + \beta_3^{(1)} y_2 + \hat{\varepsilon}$$

$$\hat{y}_2 = \beta_0^{(2)} + \beta_1^{(2)} x_1 + \beta_2^{(2)} x_2 + \beta_3^{(2)} y_1 + \hat{\varepsilon}$$

Repeated cycles are continued for a pre-specified number of rounds or until stable imputed values occur.

Necessary modifications are made to incorporate restrictions and bounds to the procedure outlined above. These restrictions are handled by fitting the models to an appropriate subset of individuals. As an example, "years passed since quitting smoking" can be asked only to former smokers. Then the fit will be restricted only to former smokers in the sample.

3.3.3. Software Packages Used

Various software packages for differing purposes are used in this study. First of all, the statistical package SPSS 15.0 version is used to compose and analyze the data sets, edit and transform the data and produce a series of tables. The software used in the imputation was IVEware, which was designed to perform imputations under SRMI approach (Raghunathan et al., 2001). Detailed introduction on this software is given in the following sub-section. Finally, ANTHRO software is used to perform anthropometric calculations required, as it was used for the calculations in TDHS-2003. Detailed information on this software is given in Section 3.2 *Anthropometric Measurements* of this thesis.

IVEware (Imputation and Variance Estimation Software) is a program for multivariate imputation of mixed, categorical and continuous variables, developed by the Survey Methodology Program of the University of Michigan. The program is based on a sequential regression algorithm that approximates the Bayesian method. The algorithm takes into account semi-continuous variables and bracketing information, and is multivariate in the sense that it conditions on the information for each case (complete or incomplete).

IVEware performs single or multiple imputations of missing values using the SRMI procedure developed by Raghunathan et al (2001). In addition to that, the software can perform a variety of descriptive and model based analyses accounting for such complex design features as clustering, stratification and weighting as well as multiple imputation analyses for both descriptive and model-based survey statistics.

The software includes four modules, which are IMPUTE, DESCRIBE, REGRESS and SASMOD.

- IMPUTE uses a SRMI approach to imputing item missing values. The module can create multiply imputed data sets.
- DESCRIBE estimates the population means, proportions, subgroup differences, contrasts and linear combinations of means and proportions. A multiple imputation analysis can be performed when there are missing values.
- REGRESS fits linear, logistic, polytomous, Poisson, Tobit and proportional hazard regression models for data resulting from a complex sample design. The repeated replication approach is used to estimate the sampling variances. A multiple imputation analysis can be performed when there are missing values.
- SASMOD allows users to take into account complex sample design features when analyzing data with several SAS procedures. A multiple imputation analysis can be performed when there are missing values. Unlike the other IVEware modules, SASMOD requires SAS.

3.4. Study Variables

To test the applicability of the proposed multiple imputation methodology to TDHS-2003 data, a new data file was prepared including the variables of interest, which are anthropometric measurements, design variables and selected auxiliary variables. This section explains the preparation of the data file used, development of the imputation model as well as definitions of key variables used in the study.

3.4.1. The Study Data File

A new data file is prepared comprising about 20 variables to be used in the study, from the Individual Data File (TRIQ41RT) of the TDHS 2003. The variables in the data file are selected to impute the anthropometric variables (height and weight) according to some criteria which are explained below.

First of all, four pillars widely used in anthropometric analysis are included in the dataset as the core variables, which are sex, age, weight and height. The reason is that, anthropometric data is available only for alive children, the variable indicating the aliveness of the children is also included in the dataset and children who are not alive on the date of interview are excluded from the data file.

In addition to these core variables, a number of auxiliary variables are also included in the dataset, which might contain potential information for imputing the missing items in anthropometric variables. Auxiliary variables, that the multiple imputation procedure will be conditioned, help to reduce the nonresponse bias and sampling variance considerably. Therefore it is generally advised to make maximal use of relevant information available, to reduce the mean squared error of prediction (Khare et al., 1993). Inclusion of auxiliary variables provides to reflect missing data uncertainty as well. Collins et al. (2001) presented a simulation to assess the potential costs and benefits of a restrictive strategy, which makes minimal use of auxiliary variables, versus an inclusive strategy, which makes liberal use of such variables. The simulation showed that the inclusive strategy is to be greatly preferred. Because prediction for missing values in each variable borrows strength from all other variables in the database (Raghunathan et al., 2001). Moreover, valid inferences in multiple imputation depends on sufficient variability in imputations, and this is ensured by the use of auxiliary variables in a conditional or *a posteriori* Bayesian sense.

Existing literature is reviewed on the factors affecting nutritional status of children, in order to find out the variables to be included in the study. There are various perspectives on the determinants of child health and nutritional status of children. Charmarbagwala et al. (2004) provide an extensive investigation of child health studies, and based on the UNICEF's framework for nutritional analysis, they summarize the main two points affecting the nutritional and health status of children in general as the following:

i. There are immediate causes (such as lack of food, low utilization of health facilities) and underlying causes which affect those immediate causes (such as family income, educational status and cultural factors which may result in gender bias in allocation of household resources) of nutrition.
ii. The determinants can be classified as child-specific (biological), household characteristics (socio-economic status) and community characteristics (service provision and cultural factors). When community level data are not available, some geographical features (urban/rural, region and cluster) are also used instead.

Charmarbagwala et al. (2004) also report the significance levels of different variables affecting the nutrition and anthropometric measures as well, included in previous studies. The variables assessed are; income (or wealth index), household size and composition (including birth order of the child), parental education, gender, location (urban vs. rural), services: sanitation, water supply and electricity, child's age, mother's age, breastfeeding, fate of previous child, health services (including place of birth), and immunization.

One of the distinctive attributions of multiple imputation procedure is that it requires imputations be conditional on the sampling design features, such as multistage stratified cluster sampling and weighting, which distinguish samples with complex designs from simple random samples. In this regard, relevant design variables are included in the data file, to ensure valid inferences.

As a result of above mentioned arguments and investigations, initially a data file comprising 26 auxiliary variables was prepared including (a) the essential information about the sample design: stratification, clustering and weighting, (b) demographic variables, (c) variables that maintain important statistical relationships between variables as well as explaining anthropometric variables (height and weight), and finally (d) variables related to the missingness. Initial data file variables are shown in Table 3.5 below. The list was constricted to 17 variables on the basis of linear regression analyses, which are described under the following section.

Table 3.6 List of Study Variables

Variable Name	% Missing	Variable Description	Variable Type
weight	5,5	weight in kg	Continuous
height	7,3	height in cm	Continuous
sex	0,0	Sex of child	Categorical
montfark	0,0	Age in months	Continuous
v001	0,0	Cluster number	Continuous
pairpsu	0,0	strata	Continuous
v002	0,0	Household number	Continuous
v003	0,0	Respondent's line number	Continuous
v005	0,0	Sample weight	Continuous
v012	0,0	Current age - respondent	Continuous
v024	0,0	Region	Categorical
v025	0,0	Type of place of residence	Categorical
v133	0,0	Education in single years	Continuous
v136	0,0	Number of household members	Continuous
bord	0,0	Birth order	Continuous
v201	0,0	Total children ever born	Continuous
v218	0,0	Number of living children	Continuous
v151	0,0	Sex of household head	Categorical
v190	0,0	Wealth index	Categorical
reswght	2,0	Respondent's weight	Continuous
reshght	1,9	Respondent's height	Continuous
breastfe	0,3	Breastfeeding	Continuous
enough	0,9	Sufficiency of antenatal care	Categorical
sizechil	0,3	Size of child at birth	Categorical
mottong	0,0	Mother tongue	Categorical
smkcigar	0,2	Smokes cigarette	Categorical
safewate	0,3	Source of drinking water grouped	Categorical
toilgr	3,4	Type of toilet facility grouped	Categorical

* Weighted calculations.

3.4.2. Definitions of Key Variables

The variables in the final study data file, in terms of their original form and several transformations to prepare the data for multiple imputation, are described in this section. First of all, it should be mentioned that dead children are excluded from the study, by using "b5" variable in the data set asking if child is alive. Children Data Set, which is constructed by using the Individual Data Set (TRIQ41RT), is used as the main data source. For some of the variables in the data set, there was information for only "usual residents"; information for other cases is taken from the Individual Data Set, in such situations.

Weight in kilograms: Weight in kilograms variable corresponds to "hw2" variable in individual data set. Since decimal points are not included in the data file, hw3 is present with one decimal place. In the study data file, hw2 is divided by 10, and missing values (999 codes) are recoded to system missing (BLANKS), to be imputed. In addition, flagged items in the data set are also recoded to system missing (BLANKS).

Height in centimetres: Height in centimetres variable corresponds to "hw3" variable in individual data set. Since decimal points are not included in the data file, hw3 is present with one decimal place. In the study data file, hw2 is divided by 10, and missing values (9999 codes) are recoded to system missing (BLANKS), to be imputed. In addition, flagged items in the data set are also recoded to system missing (BLANKS).

Sex of child: Sex of child is present as "b4" variable in the individual data set. No transformation was made on this variable.

Age in months: Sex of child is present as "hw1" variable in the individual data set. No transformation was made on this variable.

V001: Cluster number is the number identifying the sample point as used during the fieldwork. This variable may be a composite of several variables in the questionnaire.

PAIRPSU: Sample strata define the pairings or groupings of primary sampling units used in the calculation of sampling errors when using the Taylor series expansion method.

V005: Sample weight is an 8 digit variable with 6 implied decimal places. All sample weights are normalized such that the weighted number of cases is identical to the unweighted number of cases when using the full dataset with no selection. This variable is used to weight all tabulations produced using the data file.

V024: *De facto* region of residence, which can take the values (1) west, (2) south, (3) central, (4) north, and (5) east.

V025: *De facto* type of place of residence, which can take the values (1) urban and (2) rural.

Education in single years: Corresponds to "v133" in the individual data set. Indicates the years the respondent spent for education, and changes between 0 and 19 for TDHS-2003 data and has no missing values.

Number of household members: Corresponds to "v136" in the individual data set. Total number of household members is the number of usual residents plus the number of visitors who slept in the house the previous night that were listed in the household schedule.

Wealth index: Wealth index is a composite index, reflecting both the economic and the socio-economic status of households, by using numerous variables including household goods, water source, toilet facilities, income, type of dwelling and etc.

(Rutstein and Johnson, 2004). This variable corresponds to "v190" in the individual data set.

Respondent's weight: Respondent's weight corresponds to "s928" variable in the data set. Since decimal points are not included in the data file, this variable is divided by 10, and missing values (9999 codes) are recoded to system missing (BLANKS), to be imputed.

Respondent's height: Respondent's height corresponds to "s926" variable in the data set. Since decimal points are not included in the data file, this variable is divided by 10, and missing values (9999 codes) are recoded to system missing (BLANKS), to be imputed.

Size of child at birth: Size of child is a categorical variable, that can take 5 levels changing from 1 "very large" to 5 "very small". Missing values (9s) and "don't know"s are recoded as system missing (BLANK) in the new data file.

Months of breastfeeding: Months of breastfeeding variable (m5) is a continuous variable. Still breastfeeding codes are recoded to the age of children, and never breastfed codes are recoded to zero, in this variable, as well as inconsistencies (97), "don't know"s (98) and missing values (99), are recoded to system missing (BLANK).

Sufficiency of antenatal care (ANC): Sufficiency of antenatal care is a composite variable, used in further analyses constructed by Akadlı Ergöçmen et al. (2005). The categories of this variable are; 0 "No ANC", 1 "Insufficient ANC), and 2 "Sufficient ANC). For the sufficiency of ANC, a woman should visit the ANC provider in the first three months, should visit at least four times during her pregnancy and receive ANC from health personnel. Those who partially perform the above listed criteria are classified as insufficient ANC, and those who do not perform any of the criteria at all are classified as No ANC.

3.4.3. Testing the Assumptions for the Multiple Imputation

As mentioned in earlier in the study (see Section *2.3.2 Multiple Imputation*), three assumptions are required for the multiple imputation. In this section, these assumptions are investigated for the study variables. The first assumption is the model for the data values, which is generally multivariate normal for the continuous variables. The second one is a prior distribution for the parameters of the data model, which is termed as imputation model here. Finally, the nonresponse mechanism is assumed to be MAR, in multiple imputation.

Normality

Normality of dependent study variables, namely weight and height is investigated by several methods. In a normal distribution, mean, median and mode values are equal. Table 3.6 shows the mean, median and mode statistics for weight and height variables. It can be seen that for both dependent variables, three values are very close to each other, with an exception in mode of height. As a similar indicator, a distribution approximates a normal distribution when skewness and kurtosis values tend toward 0[6]. Skewness and kurtosis values of weight and height are given in Table III.7, where it can be concluded that both distributions are close to normal distribution.

Table 3.7 Basic statistics for normality tests of dependent study variables

Statistics	Weight	Height
Mean	12.94	86.58
Median	13.20	88.30
Mode	13.50	100.30
Skewness	-0.07	-0.45
Std. Error of Skewness	0.04	0.04
Kurtosis	0.06	-0.43
Std. Error of Kurtosis	0.08	0.08

[6] In a normal distribution, skewness value is 0 and kurtosis value is 3. However, statistical software packages generally give the results as kurtosis-3 (http://www.ats.ucla.edu/stat/spss/faq/kurtosis.htm).

The Imputation Model

In order to generate proper multiple imputations in a multivariate setting, a parametric model for the complete data along with a prior distribution for the parameters are specified, and values are simulated from a conditional distribution of the missing data given the observed data. Parametric model in this study is developed by linear regression analyses, since the both of the variables to be imputed, namely height and weight are continuous variables[7]. Dummy variables were created for the categorical variables in the data file. Since the amount of missing values, where none of the weight and height measurements exists, is high (7.6%), two separate linear regression models are developed in which height and weight are dependent variables in each. The reason for fitting two different regression models is to obtain covariates for both of the dependent variables, and include in the multiple imputation model as much covariates as possible. All the related variables are included in the model and a stepwise selection procedure is performed. Thus, the first model of interest when the dependent variable is weight is as follows;

Weight = -9.29 + Height * 0.25 – Size of child at birth * 0.18 + Respondent's weight * 0.01 – Female dummy * 0.23 + Education in single years * 0.02

The corresponding Model Summary Table below shows the strength of the relationship between the model and the dependent variable weight. The R Square and adjusted R square, which takes into account the number of explanatory variables in relation to the number of observations, are the same.

Table 3.8 Model Summary Table for Dependent Variable Weight

R	R Square	Adjusted R Square	Std. Error of the Estimate
0.948	0.898	0.898	1.228

[7] It is assumed that there is no type of correspondence problems among variables included in the regression.

Similarly, from the ANOVA Table which is given below, it is seen that about 90% of the variation in weight is explained by the regression model, and the model is statistically significant (sig.<0.05).

Table 3.9 ANOVA Table for Dependent Variable Weight

	Sum of Squares	F	Sig.
Regression	47742.76	6329.49	0.00
Residual	5434.25		
Total	53177.01		

The second model of interest when the dependent variable is height is as follows;

Height = 34.42 + Weight * 2.04 + Age in months * 0.36 + Wealth index * 0.26 + Respondent's height * 0.10 – Respondent's weight * 0.03 – Number of household members * 0.06 + Breastfeeding * 0.03 – East * 0.61 – South * 0.59 – North * 0.54 + Sufficiency of antenatal care * 0.19

The model summary table below shows that about 95% of the variation in height is explained by the model, which displays that the model fits the data nearly perfectly.

Table 3.10 Model Summary Table for Dependent Variable Height

R	R Square	Adjusted R Square	Std. Error of the Estimate
0.970	0.942	0.941	3.425

The ANOVA Table for Height (Table 3.10) is given below, which shows the same results, as well as the significance of the model in statistical terms (sig.<0.05).

Table 3.11 ANOVA Table for Dependent Variable Height

	Sum of Squares	F	Sig.
Regression	680258.38	5270.60	0.00
Residual	42195.76		
Total	722454.14		

Following variables were eliminated during stepwise regression procedure; current age of respondent, mother tongue, smokes cigarette, source of drinking water grouped, type of toilet facility grouped, sex of household head. Although the regression model excluded the variable "type of place of residence", it is included because of some *a priori* considerations. Collinearity diagnostics were assessed by using tolerance and variance inflation factor (VIF). There was high collinearity between number of household members, total children ever born, number of living children and birth order; consequently the last three variables were excluded from the model.

Missing Data Mechanism

A variable is MAR when missingness does not depend on the true variable of the missing variable, but it might depend on the value of other variables that are observed. The missingness mechanism in this study depends on some other variables as shown in *Missing Data on Anthropometric Measurements* section (see Section 3.2.1).

In recent debates on the plausibility of MAR, Schafer and Graham (2002) stated:

> "When missingness is beyond the researcher's control, its distribution is unknown and MAR is only an assumption. In general, there is no way to test whether MAR holds in a dataset, except by obtaining follow-up data from nonrespondents or by imposing an unverifiable model."

Furthermore, some missing data experts suggest that using missing data estimation may be better than listwise deletion approach, even though data are not MAR (Yuan and Bentler, 2000).

In this study, data are assumed as MAR, which refer that missingness on weight and height variables can depend on other variables, but it can not depend on weight and height variables themselves (e.g. because of a deficiency in height or weight of child, the respondent refused the measurement).

3.5. Imputation Procedure for the Anthropometric Variables in TDHS-2003

In this section the main features of the multiple imputation procedure implemented in the thesis is discussed. The implementation of multiple imputation in this study is based on the procedure described by Raghunathan et al. (2001), by using software IVEware as mentioned above (see Sections 3.3.1 and 3.3.2). Following this procedure, M=20 multiply imputed data sets are created with p=*4* iterations to ensure the uncorrelated imputations. The imputation procedure incorporated many predictors, including demographic and health related variables, which were included according to the results of a stepwise linear regression analysis. Small percentages of missing values in five of these predictor variables (respondent's weight, respondent's height, months of breastfeeding, sufficiency of antenatal care and size of child at birth) are also imputed.

In addition to the study variables that were incorporated into the imputation model, the study data file included the dummy variables that were used in linear regression as well as some additional auxiliary variables that the ANTHRO software required in order to calculate the anthropometric indexes. These are mainly the variables indicating the exact dates of visit (of measurement) and the date of birth of children in day, month and year detail[8]. The reason is that the anthropometric measurements, as well as ANTHRO software are very sensitive to days. Although TDHS-2003 data set include a variable as "age in months" for children, the ANTHRO software does not make use of this variable, rather it calculates the duration between the date of visit and date of birth, in days.

[8] Variables for day, months and year of visit in TDHS-2003 are; hw17, hw18 and hw19 respectively, and the variables day, month and year of birth of children are; hw16, b1 and b2 respectively.

However in TDHS-2003, data on the day of birth of children has not been collected. These unknown days of months are assigned day “15” (Rutstein and Rojas, 2003). However this imputation caused 11 children to be inconsistent since their date of visit is earlier than their date of birth. Thus these children are flagged in TDHS-2003 datasets. In order to avoid this pitfall, day variable of these children are assigned “1”, in this study.

In addition to this, problems occurred for children who are 0 month old in TDHS-2003, in the imputation process by IVEware. Therefore an additional variable is created similar to the one used in ANTHRO, which is the difference between date of visit and date of birth. Though the problem of imputation for children under age 1 month is overcame partially by using the difference variable for age in months; imputes were still problematic (height and weight measurements were too high) for some portion of children and flagged during calculation of indexes. This problem solved by imputing the children under 1 month old separately from the ones who are older than 1 month old. Creating a separate variable enabled to put bounds and restrictions for the imputation process, in order to prevent too high imputations.

Some restrictions and bounds are set for the multiple imputation process, including months of breastfeeding can not exceed the age in moths, and specified limitations for both respondent’s and child’s weight and height. The limitations for the measurements are set by considering the minimum and maximum values in the TDHS-2003 data sets and flag limitations of ANTHRO. The syntax codes document for multiple imputation application in IVEware is given in Appendix A of the study. All efforts made in order not to preclude any cases from the study were not enough for two cases, which were found to be inconsistent in terms of imputed measurements and who are older 1 month old. Imputation procedure is repeated for several times to see if the inconsistency problem can be removed for

these two cases, however all results were very similar for them. Therefore the tabulations were made by precluding these two cases.

The output file of the IVEware multiple imputation for this study is given in Appendix B. Number of imputed cases for each variable as well as basic descriptive statistics for observed part, imputed part and the combined part in each step of multiple imputation can be seen from the output. All 20 completed data sets after multiple imputation are optionally requested in one data file from the IVEware. This data file is then analysed through the formulas of Rubin (1988), given in *Multiple Imputation* section (see Section 3.3.1).

4. ANALYSES AND RESULTS

Results of multiple imputation application for the missing weight and height measurements of children under age five in TDHS-2003 are given in this section. First of all, point estimates and estimated standard errors of the study variables obtained with multiple imputation are compared with the ones associated with no imputation and single imputation. Possible bias concerned with the missingness mechanism on the study variables are also investigated, in this regard.

Imputed data is processed and analysed as is in the TDHS-2003 Main Report (HUIPS, 2004)[9] to allow comparisons. In this context, multiply imputed data is analysed according to Rubin's (1988) rules and final data file is produced. Anthropometric indexes, namely weight for age, height for age and weight for height, are calculated by using the multiply imputed data set, and compared with the ones in the TDHS-2003 Main Report (HUIPS, 2004).

In the subsequent parts of this section, these comparisons of results according to different respondent characteristics are presented. Moreover, additional characteristics of the respondents that are not presented in TDHS-2003 Main Report (HUIPS, 2004) are also tabulated and presented, in this section. All calculations are done for both before and after multiple imputation data sets, for comparisons.

4.1. Analyses of the Imputed Data

A good imputation method should accurately preserve important aspects of data distributions of the variables involved and the important relationships between them

[9] Anthropometric measurements are illustrated in Chapter 12: Infant Feeding Practices and Children's and Women's Nutritional Status of the TDHS-2003 Final Report.

(Schafer and Graham, 2002, Khare et.al., 1993). Descriptive statistics as well as histogram charts of the study variables, weight and height are presented below.

Table 4.1 Descriptive Statistics for variable "Weight", Before and After Imputation, TDHS-2003

Statistic	No imputation	Multiple imputation
Mean	12,97	12,96
Median	13,20	13,20
Mode	13,50	13,50
Skewness	-0,06	-0,09
Std. Error of Skewness	0,04	0,04
Kurtosis	0,02	-0,02
Std. Error of Kurtosis	0,08	0,08
Range	28,00	28,00
Minimum	2,20	2,20
Maximum	30,20	30,20
Total Cases	3752	3998

Table 4.1 displays the descriptive statistics (weighted) for the variable weight, both before and after multiple imputation. Though it wasn't reflected in the table, weight measurements of 253 (5.8%) children are missing or inconsistent according to actual (unweighted) results; thus couldn't be used for index calculations. It can be stated that the multiply imputed data well imitates the observed data, where mean, median and mode and exactly the same in both sides as well as the minimum and maximum values. Although the standard errors of skewness and kurtosis are preserved, kurtosis and skewness values are not fully reproduced in the imputed values. Both conclusions can be drawn by examining the Histograms of observed and multiply imputed data, in Figure 4.1 below.

Figure 4.1 Histograms Charts for Variable "Weight", Before and After Imputation, TDHS-2003

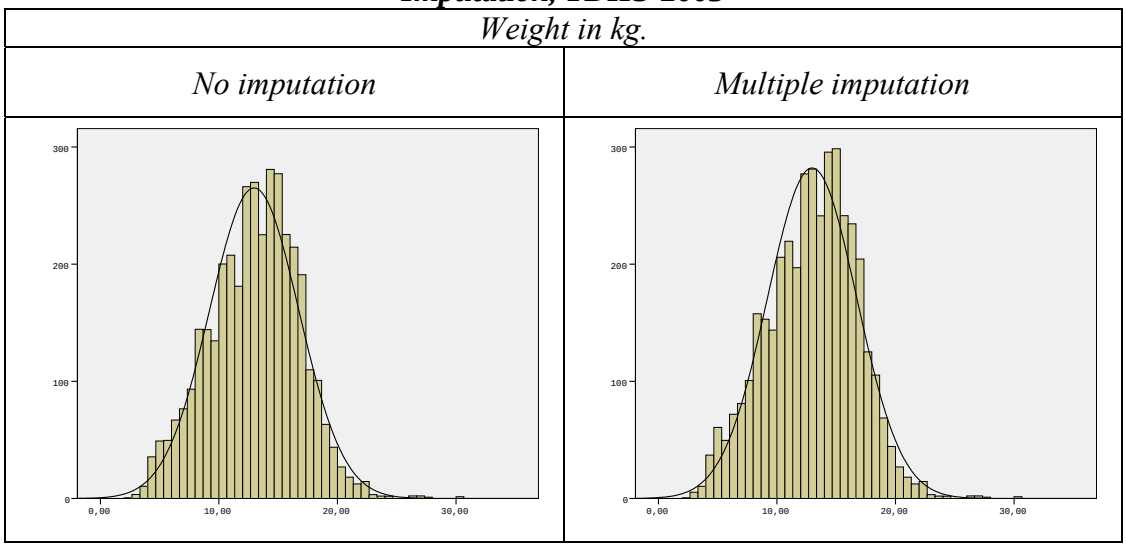


Table 4.2 displays the descriptive statistics (weighted) for variable height before and after imputation. Though it wasn't reflected in the table, height measurements of 332 (7.6%) children are missing or inconsistent according to actual (unweighted) results; thus couldn't be used for index calculations. Similar to weight, the multiply imputed data well imitates the observed data, which can also be seen from Figure 4.2 Histograms below. In general terms, multiply imputed values are consistent with the observed data, for both of the study variables.

Table 4.2 Descriptive Statistics for variable "Height", Before and After Imputation

Statistic	No imputation	Multiple imputation
Mean	86,78	86,76
Median	88,40	88,40
Mode	100,30	100,30
Skewness	-0,42	-0,45
Std. Error of Skewness	0,04	0,04
Kurtosis	-0,50	-0,44
Std. Error of Kurtosis	0,08	0,08
Range	72,90	73,00
Minimum	49,10	49,00
Maximum	122,00	122,00
Total Cases	3678	3998

Figure 4.2 Histograms Charts for Variable "Height", Before and After Imputation, TDHS-2003

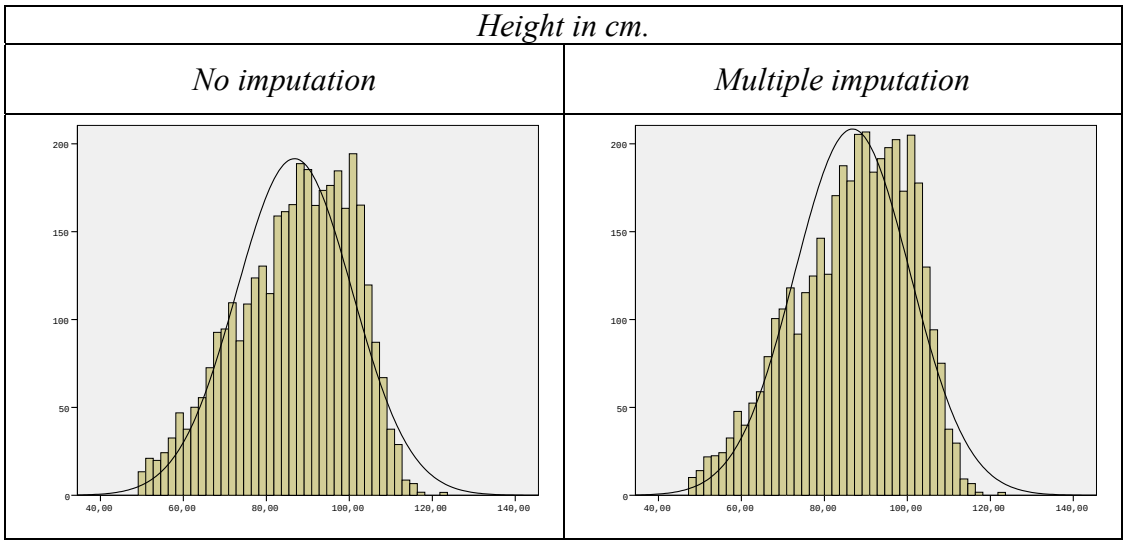


Table 4.3 displays point estimates and estimated standard errors for the seven study variables imputed, based on TDHS-2003 with no imputation, with single imputation and with multiple imputation. Single imputation estimates are computed by using the first imputed complete dataset of the 20 multiple imputations. Table 4.3 also displays the ratios of estimated standard errors based on no imputation as well as single imputation to the ones based on multiple imputation.

Table 4.3 Point Estimates, Estimated Standard Errors (SEs) and Approximate Fractions of Missing Information (RM) for the Imputed Variables in TDHS-2003 Data

	No imputation (NI)		Single imputation (SI)		Multiple imputation (MI)			Ratio of estimated SEs	Ratio of estimated SEs
Imputed variables	Point estimate	Estimated SE	Point estimate	Estimated SE	Point estimate	Estimated SE	*R*	NI/MI	SI/MI
Weight in kg	12,97	0,07	12,95	0,07	12,96	0,07	1,13	0,996	0,994
Height in cm	86,78	0,26	86,76	0,25	86,76	0,25	0,84	1,019	1,002
Respondent's weight	64,78	0,27	64,79	0,27	64,78	0,27	2,40	1,001	0,991
Respondent's height	156,59	0,15	156,57	0,15	156,60	0,15	1,33	1,003	0,987
Breastfeeding	11,15	0,17	11,17	0,17	11,16	0,17	0,31	1,000	0,999
Sufficiency of ANC	1,20	0,02	1,20	0,02	1,20	0,02	0,30	1,002	1,002
Size of child at birth	3,22	0,02	3,22	0,02	3,22	0,02	0,52	1,000	0,997

Point estimates in no imputation, single imputation and multiple imputation are very close to each other in general, as seen from the Table. This result is reasonable for the variables other than weight and height because of lower percentages of associated missing data (i.e.<5%). For the study variables weight and height, however, a relative difference between point estimates based on no imputation and estimates based on imputations, is seen, which may refer to a possible bias in complete cases.

Mechanism of missingness on weight and height was discussed in Section 3.2.1. *Missing Data on Anthropometric Measurements.* Four variables which were found to be in relation with weight and height are examined here, as an attempt to explain the difference in point estimates of weight and height based on no imputation and based on imputation.

Figure 4.3 illustrates that there are remarkable differences in the percentages of missing data in weight and height variables among different categories of respondent characteristics. As discussed earlier, "0 month" in A*ge in months* variable is somewhat problematic, this is also reflected to the Figure 4.3 below. In both weight and height, percentage of missing data is high in "0-6 months" age group. In height, "24-35 months" age group is also apparently high in terms of nonresponse. The slight difference in point estimates of weight and height mentioned above, may be a result of the imputation of "0 month" age group.

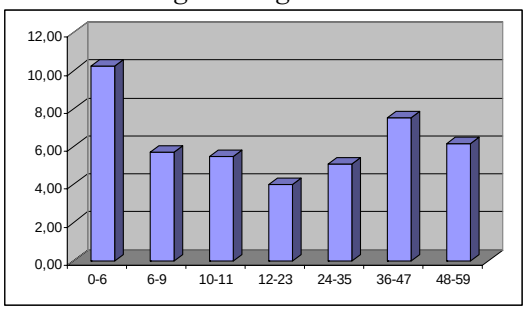
In addition, Figure 4.3 displays that for both variables of interest, percentage of missing data is particularly high in the last education category, which is "high school and higher". Higher percentages of missing data are also observed in higher wealth index categories as well as female household head.

In theory, the estimated standard errors with no imputation are expected to be higher than those with multiple imputation, thus the multiple imputation results in more precise point estimates than complete case analysis. Similarly, estimated standard errors with single imputation are expected to be lower than those with multiple imputation; reflecting that the single imputation analyses do not account for additional uncertainty due to imputation. However, these are not the cases for all variables as seen in the last two columns of Table 4.1. These results may be related with the biases in anthropometric measurements; because as Schenker et.al. (2006) state, estimated standard errors for complete data can be biased if the missing data mechanism is not MCAR.

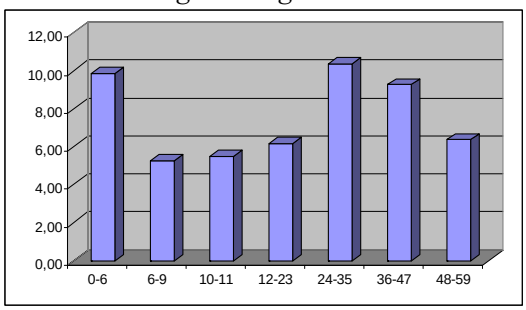
Approximate fractions of missing information in Table 4.1 are all around 2% and are much smaller than the percentages of missing data in TDHS-2003 data set. As introduced in literature (Schafer and Graham, 2002, Schenker et.al., 2006), the fraction of missing information tends to be smaller than the percentage of missing data in practice, because the fraction of missing information accounts for the information incorporated into the imputation model that is predictive of the variable subject to nonresponse, additionally. As well as that, the fraction of missing information is specific to the amount of data being imputed. As a result, fraction of missing information both reflects the amount of imputed data and how well the imputation model predicts values; where smaller fractions are desirable. Based on the small fractions in Table 4.3, it can be concluded that the imputation model used well predicts the dependent variables in the study.

Figure 4.3 Distribution of Percentage of Missing Values in Weight and Height according to Age in Months, Education in Categories, Wealth Index and Sex of Household Head in TDHS-2003 Data

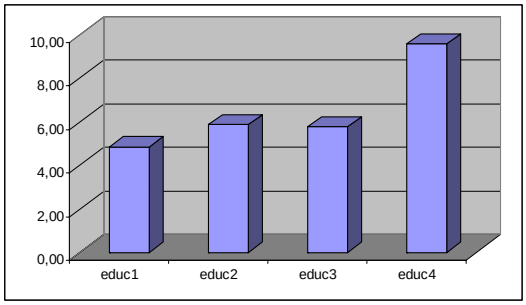
% Weight vs. Age in Months



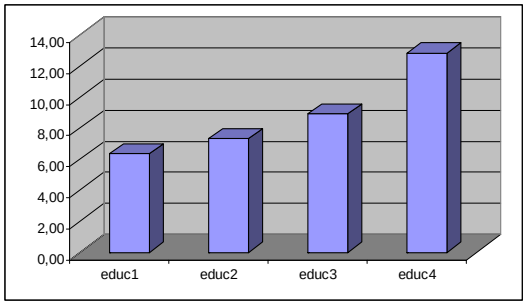
% Height vs. Age in Months



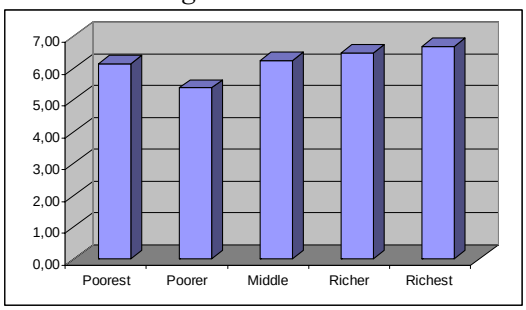
*% Weight vs. Education**



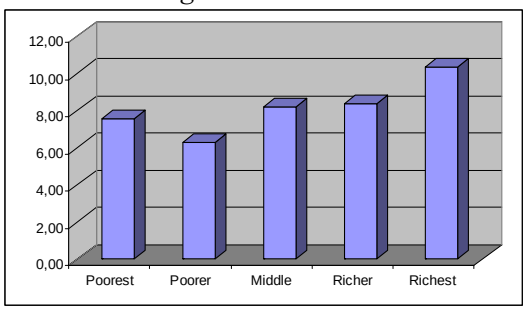
*% Height vs. Education**



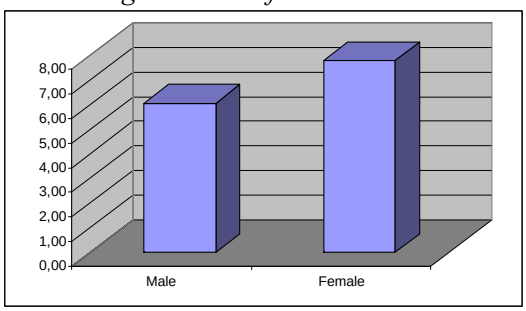
% Weight vs. Wealth Index



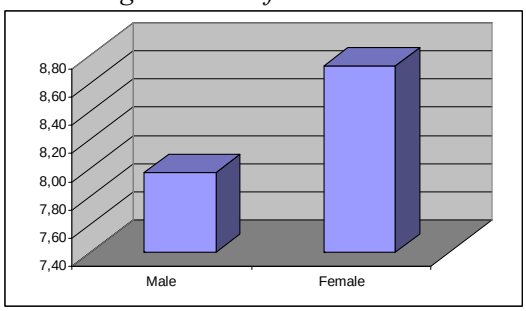
% Height vs. Wealth Index



% Weight vs. Sex of Household Head



% Height vs. Sex of Household Head



* Education categories are coded as follows; educ1=No education/primary incomplete, educ2= First level primary, educ3=Second level primary, educ4=High school and higher.

4.2. Comparison of Multiple Imputation Results Before and After Multiple Imputation for the Anthropometric Indexes

In this section, anthropometric indexes, namely height for age, weight for height and weight for age, for children under age 5 that are obtained before and after multiple imputation are compared. As discussed above, distributions of the study variables are preserved in general terms. In this section, the relationships with some selected background characteristics of the respondents are investigated.

Tables in this section are given for children in the period 0-59 months preceding the survey. Indexes are expressed in terms of the number of standard deviation (SD) units from the median of the NCHS/CDC/WHO international reference population. Children are classified as malnourished if their z-scores are below minus two or minus three standard deviations (-2 SD or – 3 SD) from the median of the reference population. In addition, in all tables given below, figures of -2SD include children who are below -3 SD. TDHS-2003 findings, as they are in the Main Report, are given as the upper (before imputation) parts of almost all tables in this section.

Table 4.4 displays the percentages of children who are less than two standard deviations below the reference population, in terms of three anthropometric indexes height for age, weight for height and weight for age. The extent to which the percentage of children falling into these categories exceeds 2.3%, which is the expected percentage in a well nourished population, shows the level of specific aspects of malnutrition in that population (Rutstein and Rojas, 2003). Definitions for anthropometric terms are given in *3.2 Anthropometric Measurements* section of this thesis. According to TDHS-2003 findings, 12.2% of the children are stunted, while 0.7% are wasted and 3.9% are underweight in Turkey. The percentages of children in these malnutrition categories decreased after multiple imputation, where it is found that 11.4% of the children are stunted, where 0.6% are wasted and 3.5% are

underweight. A similar decrease is observed for the percentage of children under -3 SD from the median of the reference population for all indexes.

Table 4.4 ***Comparison of No Imputation and Multiple Imputation Model Results for Nutritional Status of Children by Child's Age***

No imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Child's age (in months)							
0-6	0,3	2,2	0,4	1,2	0,2	0,8	334
6-9	3,0	5,6	0,0	0,8	0,0	1,7	247
10-11	2,8	10,8	0,4	1,5	1,9	5,7	103
12-23	1,4	12,4	0,4	0,8	0,5	2,9	702
24-35	3,5	12,2	0,7	1,0	1,3	5,2	755
36-47	6,0	15,4	0,0	0,3	0,3	5,1	750
48-59	5,3	15,4	0,2	0,3	0,6	4,1	777
Total	3,6	12,2	0,3	0,7	0,6	3,9	3668
Multiple imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Child's age (in months)							
0-6	0,3	2,0	0,6	1,3	0,2	0,7	370
6-9	2,0	5,3	0,0	0,8	0,0	1,6	261
10-11	2,7	10,2	0,4	1,5	1,8	5,4	109
12-23	1,2	11,8	0,3	0,7	0,5	2,6	749
24-35	3,2	11,3	0,6	0,9	1,2	4,6	845
36-47	5,7	14,6	0,0	0,3	0,3	4,7	829
48-59	5,0	14,2	0,2	0,2	0,5	3,7	833
Total	3,4	11,4	0,3	0,6	0,6	3,5	3996

Table 4.5 shows the percentage of children under five years, classified as malnourished according to three anthropometric indexes by child's age in single months, for both before and after imputation. The decrease in height for age figures between no imputation and multiple imputation is clear, when the figure is examined.

Table 4.5 Comparison of No Imputation and Multiple Imputation Model Results for Nutritional Status of Children by Child's Age

	No imputation				*Multiple imputation*			
	Height for age	Weight for height	Weight for age		Height for age	Weight for height	Weight for age	
Age in months	Percentage below -2SD	Percentage below -2SD	Percentage below -2SD	Number of children	Percentage below -2SD	Percentage below -2SD	Percentage below -2SD	Number of children
<2	0,8	4,4	0,8	74	0,6	4,4	0,6	94
2-3	3,8	0,0	0,0	122	3,6	0,0	0,0	128
4-5	1,6	0,6	1,4	138	1,5	0,5	1,4	148
6-7	7,4	0,2	2,5	135	7,0	0,2	2,4	143
8-9	3,3	1,5	0,8	112	3,2	1,4	0,8	118
10-11	10,8	1,5	5,7	103	10,2	1,5	5,4	109
12-13	10,9	1,6	3,6	132	10,1	1,6	3,4	138
14-15	11,3	0,0	3,6	101	10,8	0,0	3,4	106
16-17	11,2	0,4	2,2	127	11,5	0,4	2,0	136
18-19	10,9	0,7	2,4	123	10,6	0,6	2,2	131
20-21	13,3	1,2	1,2	95	11,8	1,1	1,1	104
22-23	17,0	0,6	3,9	124	15,8	0,6	3,2	134
24-25	8,2	3,7	8,4	97	7,3	3,3	7,6	108
26-27	8,6	0,0	5,5	114	8,7	0,0	5,1	122
28-29	9,9	1,6	4,8	144	8,0	1,4	4,2	166
30-31	12,1	0,0	5,4	127	11,0	0,0	4,0	144
32-33	16,0	1,2	3,3	137	15,3	1,1	2,6	151
34-35	17,0	0,0	4,7	137	16,2	0,0	4,9	154
36-37	13,8	1,2	5,2	119	12,5	1,0	4,8	141
38-39	12,6	0,0	2,4	112	11,7	0,0	2,1	129
40-41	9,2	0,0	2,6	139	9,9	0,0	2,6	151
42-43	19,2	0,0	6,3	141	18,3	0,0	6,0	148
44-45	15,2	0,0	5,5	103	14,3	0,0	4,9	116
46-47	21,8	0,8	8,3	135	20,4	0,7	7,8	144
48-49	21,0	0,0	5,0	112	19,3	0,0	4,2	124
50-51	14,7	0,2	2,9	132	13,8	0,0	2,7	141
52-53	11,0	0,2	4,7	135	9,2	0,2	4,3	146
54-55	14,8	0,0	1,6	143	13,9	0,0	1,5	149
56-57	13,3	0,8	4,8	135	12,5	0,8	4,1	143
58+	19,2	0,7	6,4	120	17,7	0,0	5,9	130
Total	12,2	0,7	3,9	3668	11,4	0,6	3,5	3996

Figure 4.4 illustrates the change in the percentages of nutritional status of children by age, for both before and after imputation status. Plotted values in the figure are smoothed by a five month moving average, as is in the TDHS-2003 Main Report. It is important to emphasize here that the pattern didn't change after the multiple imputation application; however the lines are shifted below, in addition to the overall decrease in percentages at all ages. This shift is also very clear for the stunted line.

Figure 4.4 Comparison of No Imputation and Multiple Imputation Model Results for Nutritional Status of Children by Age

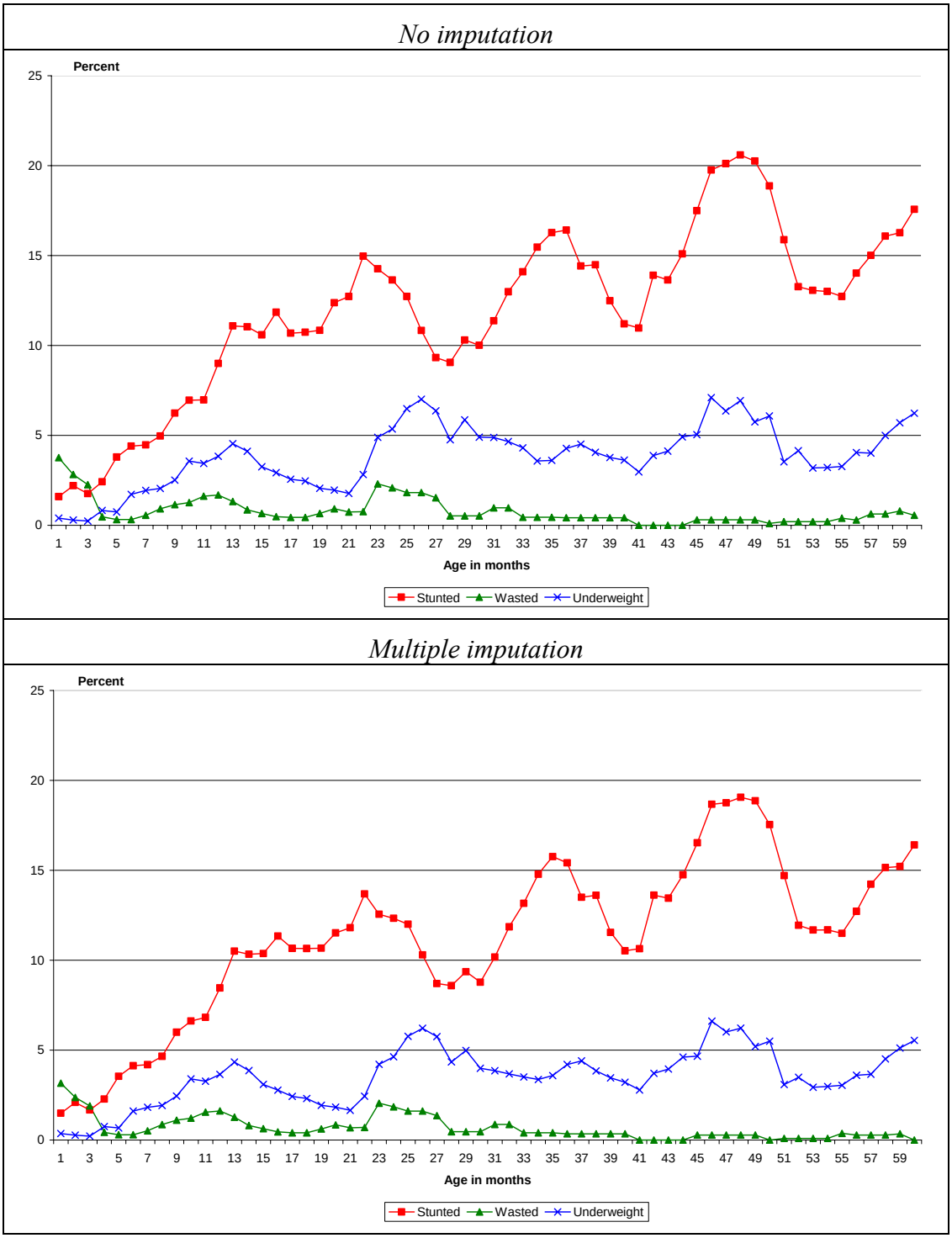


Figure 4.5 presents a clearer comparison for before and multiple imputation, for the percent of children who are defined as "stunted" according to NCHS/WHO/CDC standards. Similarly, Figure 4.6 and Figure 4.7 present clearer comparisons for

before and multiple imputation, for the percent of children who are defined as "wasted" and "underweight" respectively, according to NCHS/WHO/CDC standards.

Figure 4.5 Comparison of No Imputation and Multiple Imputation Model Results for Stunting Status Children by Age

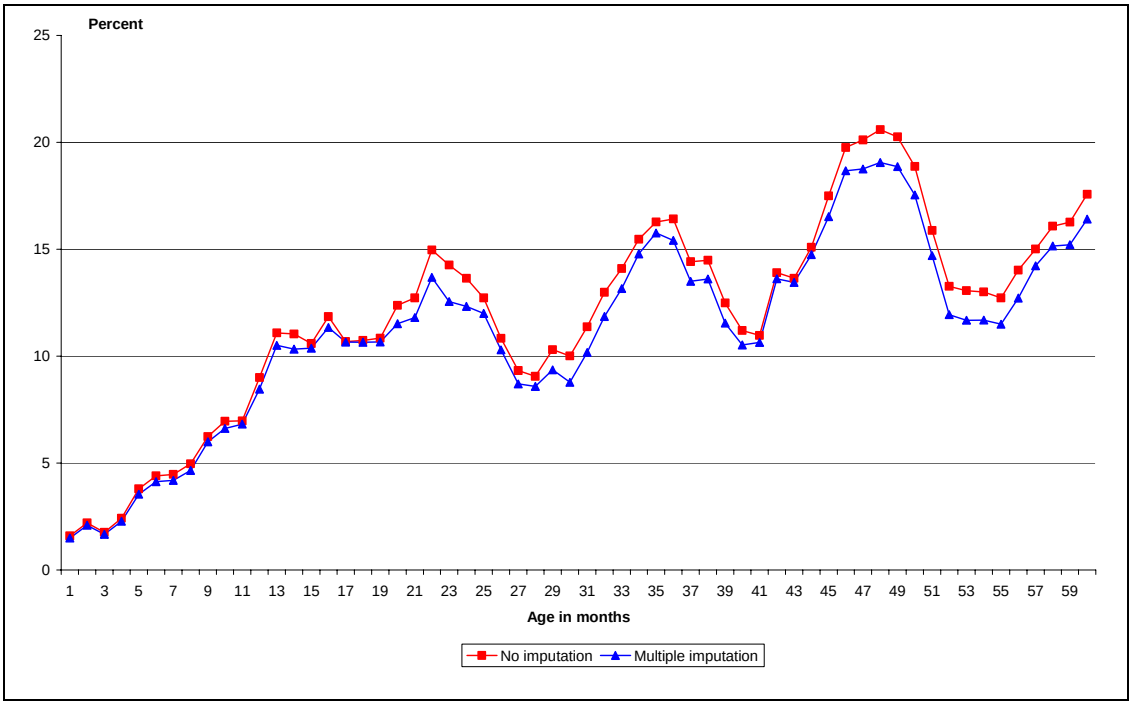


Figure 4.6 Comparison of No Imputation and Multiple Imputation Model Results for Wasting Status Children by Age

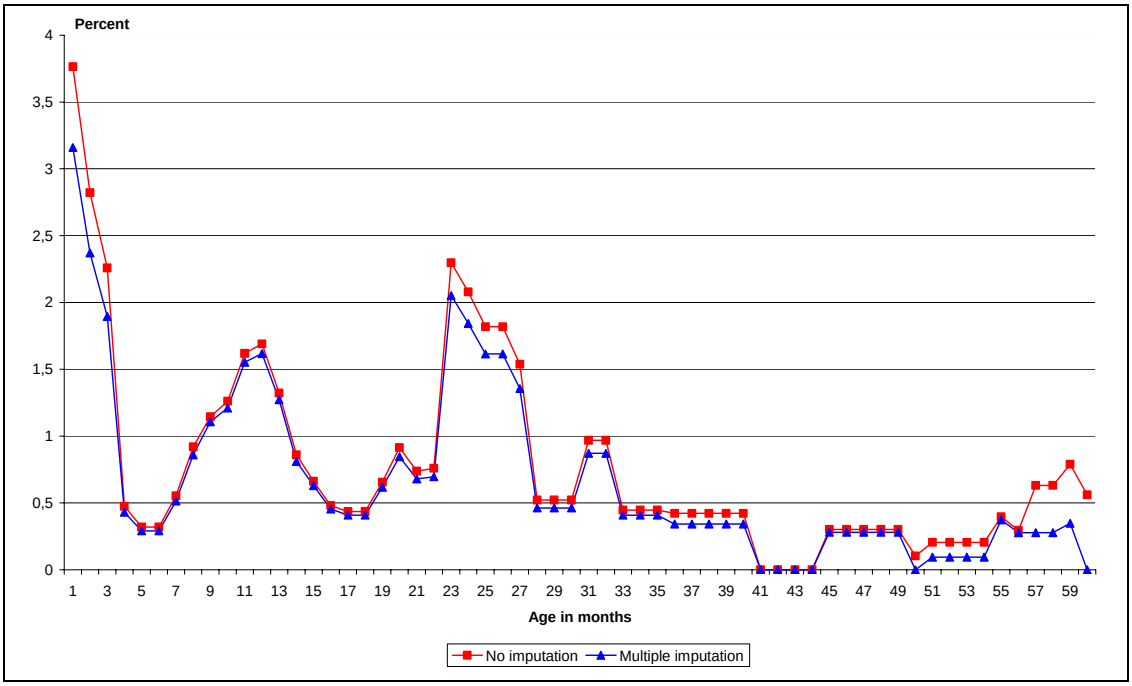


Figure 4.7 Comparison of No Imputation and Multiple Imputation Model Results for Underweight Status Children by Age

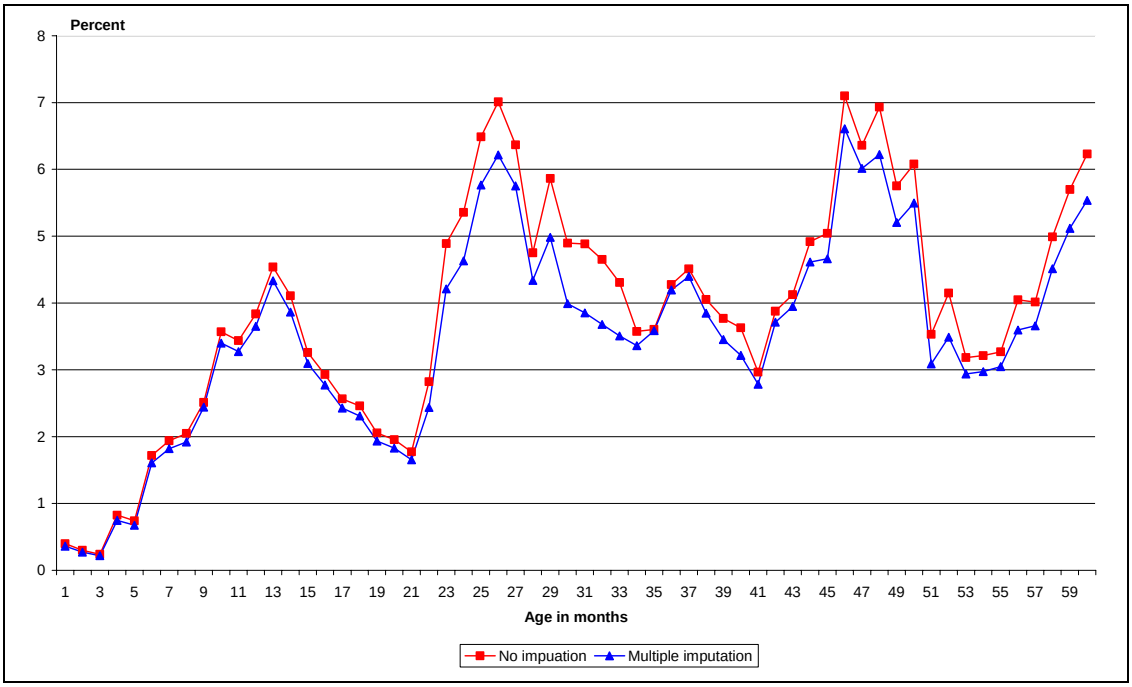


Table 4.6 Comparison of No Imputation and Multiple Imputation Model Results for Nutritional Status of Children by Sex of Child

	No imputation						
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Sex of child							
Male	2,9	10,9	0,4	1,0	0,6	3,2	1890
Female	4,5	13,6	0,1	0,4	0,7	4,7	1778
Total	3,6	12,2	0,3	0,7	0,6	3,9	3668
	Multiple imputation						
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Sex of child							
Male	2,6	10,0	0,4	0,9	0,5	2,8	2061
Female	4,2	12,8	0,1	0,3	0,6	4,3	1935
Total	3,4	11,4	0,3	0,6	0,6	3,5	3996

In Table 4.6 nutritional status of children under age 5 by sex of child is presented for before and after imputation application. Percentage of children below -2 SD for

height for age decreased at 7% and 5% levels for male and female children respectively, after multiple imputation. Slight decreases in weight for height and weight for age are also observed.

Table 4.7 Comparison of No Imputation and Multiple Imputation Model Results for Nutritional Status of Children by Birth Order

No imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Birth order							
1	2,0	7,2	0,3	0,5	0,3	2,1	1225
2-3	2,4	10,3	0,2	0,7	0,5	3,3	1614
4-5	7,4	21,1	0,8	1,7	2,1	8,2	468
6+	9,7	26,0	0,0	0,4	1,0	7,1	361
Total	3,6	12,2	0,3	0,7	0,6	3,9	3668
Multiple imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Birth order							
1	1,7	6,5	0,4	0,5	0,2	1,8	1364
2-3	2,3	9,7	0,2	0,6	0,4	3,0	1742
4-5	6,8	19,5	0,7	1,5	1,9	7,5	508
6+	9,7	25,8	0,0	0,2	0,9	7,0	382
Total	3,4	11,4	0,3	0,6	0,6	3,5	3996

In Table 4.7, nutritional status of children is displayed by the birth order, for before and after imputation. The percentage below -2 SD in height for age for 6+ births is dramatically high when compared to other birth orders.

Differences between urban-rural residences as well as regions are seen in Table 4.8, in terms of the percentages under certain levels according to TDHS-2003 results. The decrease in percentages before and after multiple imputation is at 7% levels in general for the percentage of children under -2 SD for all indexes; this level is higher in South and lower in East regions.

Table 4.8 Comparison of No Imputation and Multiple Imputation Model Results for Nutritional Status of Children by Residence and Region

	Height for age		Weight for height		Weight for age		
No imputation							
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Residence							
Urban	2,6	9,0	0,3	0,7	0,6	2,8	2414
Rural	5,6	18,4	0,3	0,8	0,8	5,9	1254
Region							
West	0,6	5,5	0,5	0,7	0,5	1,9	1186
South	2,7	10,4	0,2	0,4	0,2	2,8	499
Central	2,6	9,5	0,3	0,8	0,8	2,9	727
North	3,7	13,0	0,2	0,7	0,0	2,2	218
East	8,3	22,5	0,1	0,8	1,1	7,7	1038
Total	3,6	12,2	0,3	0,7	0,6	3,9	3668
Multiple imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Residence							
Urban	2,4	8,3	0,3	0,6	0,5	2,5	2649
Rural	5,3	17,4	0,3	0,7	0,7	5,5	1347
Region							
West	0,5	5,0	0,4	0,6	0,5	1,7	1304
South	2,5	9,7	0,2	0,4	0,2	2,4	542
Central	2,1	8,7	0,2	0,8	0,7	2,7	789
North	3,4	11,7	0,2	0,6	0,0	2,0	242
East	7,9	21,5	0,2	0,7	0,9	7,2	1118
Total	3,4	11,4	0,3	0,6	0,6	3,5	3996

Table 4.9 shows the percentage of children classified as malnourished according to anthropometric indexes by NUTS1 Regions of Turkey. A striking issue to mention is the increase in the percentage of children in Northeast Anatolia NUTS1 Region, under -2 SD for height for age and weight for height indexes. For the remaining NUTS1 Regions, general pattern of decrease is seen from the table below.

Table 4.9 Comparison of No Imputation and Multiple Imputation Model Results for Nutritional Status of Children by NUTS1 Regions

No imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
NUTS 1 Region							
Istanbul	0,9	6,1	0,3	0,7	0,4	1,3	572
West Marmara	1,0	7,3	0,7	0,7	0,7	6,3	113
Aegean	1,6	6,6	0,0	0,8	0,8	1,2	346
East Marmara	0,4	3,4	1,1	1,6	1,1	3,0	284
West Anatolia	2,2	9,8	0,4	0,4	0,4	2,7	311
Mediterranean	2,7	10,4	0,2	0,4	0,2	2,8	499
Central Anatolia	1,6	9,6	0,4	0,4	0,8	2,4	204
West Black Sea	3,1	9,1	0,0	0,6	0,0	3,0	182
East Black Sea	4,3	16,9	0,3	0,3	0,0	2,3	118
Northeast Anatolia	6,7	16,8	0,2	1,3	0,9	6,7	166
Central East Anatolia	10,1	26,6	0,0	0,3	1,3	9,6	280
Southeast Anatolia	8,0	22,1	0,2	0,9	1,0	7,1	592
Total	3,6	12,2	0,3	0,7	0,6	3,9	3668
Multiple imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
NUTS 1 Region							
Istanbul	0,9	5,7	0,3	0,6	0,3	1,0	624
West Marmara	1,0	6,9	0,7	0,7	0,7	6,0	119
Aegean	1,4	5,7	0,0	0,7	0,7	1,1	383
East Marmara	0,4	2,8	1,0	1,5	1,0	2,7	317
West Anatolia	1,3	8,8	0,3	0,3	0,3	2,4	345
Mediterranean	2,5	9,7	0,2	0,4	0,2	2,4	542
Central Anatolia	1,4	8,8	0,4	0,4	0,7	2,2	221
West Black Sea	2,9	9,0	0,0	0,6	0,0	2,8	193
East Black Sea	3,8	15,0	0,3	0,3	0,0	2,1	133
Northeast Anatolia	6,9	18,0	0,6	1,5	0,8	6,5	176
Central East Anatolia	9,6	24,8	0,0	0,0	1,3	9,0	300
Southeast Anatolia	7,4	21,0	0,2	0,8	0,8	6,6	642
Total	3,4	11,4	0,3	0,6	0,6	3,5	3996

Table 4.10 shows the percentage of children classified as malnourished according to anthropometric indexes by birth interval, where higher levels of malnutrition is seen at short birth interval levels, in both before and after imputation. The decreasing pattern between before and after imputation is seen clearly in first births than other categories of birth interval.

Table 4.10 Comparison of No Imputation and Multiple Imputation Model Results for Nutritional Status of Children by Birth Interval

No imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Birth interval							
First birth	2,0	7,2	0,3	0,5	0,3	2,1	1239
Under 24 months	7,7	21,0	0,5	1,4	2,0	7,0	646
24-47 months	4,9	16,0	0,0	0,7	0,3	5,2	888
48+ months	1,7	8,9	0,3	0,5	0,6	2,8	896
Total	3,6	12,2	0,3	0,7	0,6	3,9	3668
Multiple imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Birth interval							
First birth	1,7	6,5	0,4	0,5	0,2	1,8	1379
Under 24 months	7,4	20,5	0,5	1,3	1,9	6,7	686
24-47 months	4,6	15,0	0,0	0,6	0,2	4,7	968
48+ months	1,6	8,4	0,3	0,4	0,5	2,6	963
Total	3,4	11,4	0,3	0,6	0,6	3,5	3996

Finally Table 4.11 presents the percentage of children classified as malnourished according to anthropometric indexes by education of the respondent, for before and after imputation. As being one of the characteristics affecting nonresponse mechanism of weight and height variables, remarkable differences in education categories are seen in the table below.

Table 4.11 Comparison of No Imputation and Multiple Imputation Model Results for Nutritional Status of Children by Education

	Height for age		Weight for height		Weight for age		
No imputation							
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Education							
No education/Primary incomplete	9,1	25,3	0,1	1,0	1,1	8,3	975
First level primary	2,1	9,0	0,3	0,6	0,6	2,7	1895
Second level primary	1,7	5,6	0,8	0,8	0,0	1,8	275
High school and higher	0,2	2,9	0,2	0,5	0,2	0,9	524
Total	3,6	12,2	0,3	0,7	0,6	3,9	3668
Multiple imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Education							
No education/Primary incomplete	8,7	24,4	0,1	0,9	1,0	7,8	1044
First level primary	1,9	8,3	0,3	0,6	0,6	2,5	2051
Second level primary	1,5	5,2	0,7	0,7	0,0	1,3	301
High school and higher	0,2	2,5	0,2	0,4	0,2	0,8	599
Total	3,4	11,4	0,3	0,6	0,6	3,5	3996

In the following section, anthropometric indexes are displayed for more background characteristics of the respondents, in addition to the ones given in TDHS-2003 Main Report (2004). These characteristics are sufficiency of antenatal care, wealth index, mother tongue and size of child at birth. As seen below, there are missing data on some of these additional characteristics. In obtaining relationships between two or more variables, the available software conducts a listwise deletion where items are taken into account only if they have values on both variables. As a result of this, in the "No imputation" part of tables, the total number of children is lesser than the ones displayed above. These variables are also imputed since they were in the context of imputation model of the study. Hence, "Multiple imputation" part of the tables show the total number of children after imputation. The tables in this section

are given for only displaying the relationship of anthropometric measurements with more background characteristics. Statistical concern is not considered such as low number of observations for categories.

In Table 4.12 below, percentage of children classified as malnourished according to anthropometric indexes by sufficiency of antenatal care is shown for before and after imputation. The table displays the role of antenatal care in nutrition status of children, where one in every four child is malnourished among children who have taken no antenatal care, though figures are lower after imputation. a striking issue in that table is that the while analysis results are given for 3668 children for anthropometric indexes in general, this number falls to 3634 children in this table, because of additional missing data on sufficiency of antenatal care variable.

Table 4.12 Comparison of No Imputation and Multiple Imputation Model Results for Nutritional Status of Children by Sufficiency of Antenatal Care

No imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Sufficiency of antenatal care							
No ANC	8,7	25,3	0,1	1,1	1,3	8,1	839
Unsufficient ANC	3,2	11,6	0,6	1,0	0,9	4,2	1269
Sufficient ANC	1,2	5,5	0,1	0,3	0,1	1,3	1526
Total	3,7	12,2	0,3	0,7	0,7	3,9	3634
Multiple imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Sufficiency of antenatal care							
No ANC	8,3	24,3	0,1	0,9	1,1	7,4	903
Unsufficient ANC	3,0	11,1	0,6	1,0	0,9	3,9	1396
Sufficient ANC	1,0	4,8	0,1	0,2	0,1	1,2	1696
Total	3,4	11,4	0,3	0,6	0,6	3,5	3996

In Table 4.13, it can be clearly seen that there are major differences among the wealth index categories in nutritional status of children, both for before and after imputations. However, there is no considerable difference between the levels of decrease among categories between before and after imputation results. Table 4.14 shows the comparison of no imputation and multiple imputation model results for nutritional status of children by mother tongue in categories.

Table 4.13 Comparison of No Imputation and Multiple Imputation Model Results for Nutritional Status of Children by Wealth Index

No imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Wealth index							
Poorest	7,7	25,0	0,4	1,2	1,2	7,8	906
Poorer	4,9	13,2	0,0	0,3	0,7	5,1	788
Middle	2,5	10,1	0,7	1,1	0,9	2,4	697
Richer	0,8	3,9	0,2	0,3	0,2	1,0	736
Richest	0,4	3,4	0,1	0,4	0,1	1,3	542
Total	3,6	12,2	0,3	0,7	0,6	3,9	3668
Multiple imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Wealth index							
Poorest	7,2	23,6	0,4	1,1	1,1	7,4	982
Poorer	4,4	12,8	0,1	0,3	0,6	4,6	847
Middle	2,3	9,1	0,7	1,0	0,8	2,2	759
Richer	0,7	3,5	0,2	0,3	0,2	0,8	802
Richest	0,3	3,0	0,1	0,3	0,1	1,0	605
Total	3,4	11,4	0,3	0,6	0,6	3,5	3996

Table 4.14 Comparison of No Imputation and Multiple Imputation Model Results for Nutritional Status of Children by Mother Tongue

No imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Mother tongue							
Turkish	1,6	7,8	0,4	0,7	0,5	2,3	2584
Kurdish	8,8	23,4	0,1	0,8	1,0	7,8	942
Other	6,0	17,7	0,0	0,7	0,4	6,0	143
Total	3,6	12,2	0,3	0,7	0,6	3,9	3668
Multiple imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Mother tongue							
Turkish	1,4	7,2	0,4	0,6	0,5	2,2	2823
Kurdish	8,3	22,2	0,1	0,7	0,9	7,2	1021
Other	5,7	16,7	0,0	0,7	0,4	4,6	152
Total	3,4	11,4	0,3	0,6	0,6	3,5	3996

As seen in Table 4.15, size of child at birth is one of the variables that has missing values in the TDHS-2003 data set. The general shift downwards is also observed for the size of child at birth, before and after imputation application.

Table 4.15 Comparison of No Imputation and Multiple Imputation Model Results for Nutritional Status of Children by Size of Child at Birth

No imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Size of child at birth							
Very large	1,7	6,3	0,0	0,0	0,0	0,0	47
Larger than average	2,5	5,9	0,0	0,1	0,1	2,0	547
Average	2,6	9,9	0,3	0,6	0,5	2,7	2031
Smaller than average	4,8	17,7	0,4	1,2	0,8	5,8	614
Very small	8,8	24,7	0,5	1,5	2,0	10,0	417
Total	3,7	12,2	0,3	0,7	0,7	3,9	3656
Multiple imputation							
	Height for age		Weight for height		Weight for age		
	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Percentage below -3SD	Percentage below -2SD	Number of Children
Size of child at birth							
Very large	1,5	5,7	0,0	0,0	0,0	0,0	52
Larger than average	2,4	5,6	0,0	0,1	0,1	1,9	595
Average	2,4	9,1	0,3	0,5	0,4	2,4	2222
Smaller than average	4,1	16,4	0,3	1,1	0,7	5,1	679
Very small	8,3	23,4	0,5	1,3	1,9	9,3	448
Total	3,35	11,40	0,29	0,63	0,58	3,53	3996

On the other hand, TDHS-2003 includes other tables regarding the nutritional status of women. Respondent's weight and respondent's height variables were also in the context of the imputation model of the study. However, it should be noted that, the variables other than weight and height of children under age five were handled as independent variables in the study. For example, if respondent's weight and respondent's height were selected as the dependent variables, they would have different covariates to explain them. In addition to that, the imputation model of the study was designed on the basis of children as units, thus data for a woman could be imputed twice or more, if she had more than one child, according to the imputation model of the study. Consequently, providing estimations or relationships for the dependent variables of the imputation model may be somewhat unfavourable.

5. CONCLUSION AND DISCUSSION

The main objective of this study was to apply a multiple imputation model applying the anthropometric measurements for children under age five, in TDHS-2003 data set. Multiple imputations are created through SRMI approach and using IVEware software, which can handle complex sample structures and different types of data.

This study is expected to contribute to the multiple imputation literature in Turkey in both technical and practical terms. Technical contribution covers the literature that is yet very limited in the country. Thus the issue is not very well known by the applied researchers in general as well. In practical terms, this study is one of the first attempts in application of multiple imputation to a nationally representative data set in Turkey.

The study is expected to contribute to the DHS imputation literature as well. Worldwide implemented DHS' has a sound imputation methodology for only six key dates, and has its own constraints and shortcomings, as documented by Croft (1991). Although not reported before, this study revealed that a number of cases are precluded from the analyses in calculation of anthropometric indexes due to the shortcomings in imputation method of DHS of imputing *day of birth* as "15" for all children, regardless of the date of visit. In this context, collection of day of birth information for all children in Turkey can be a recommendation presented by this study, since the calculation of anthropometric measurements is very sensitive to the day of birth information.

According to the results of this study, multiply imputed data well imitated the observed data in terms of distribution for both of the study variables, weight and height. Although the central tendency values (mean, median and mode) are very

similar to each other before and after imputation; minor differences observed in point estimates of both of the study variables, which gave an idea about a possible bias in the observed data. In addition, it has been seen that there is a remarkable differentiation in the percentages of missing data according to different categories of respondent characteristics.

Proportion of missing data in measurements for especially younger children, higher education levels of respondents, higher levels of wealth index and female household headed households were relatively high. Moreover, estimated standard errors with multiple imputation are not lower than those with no imputation for all of the study variables, according to the analyses results, which may also be related to a possible bias due to nonresponse associated with the study variables. All in all, though it cannot be asserted that there is an exact bias among observed and unobserved values, since the real values of the population represented cannot be known; it can be expressed that multiple imputation adjusted for the bias due to nonresponse among these values.

When the relationships between anthropometric indexes based on not imputed and multiply imputed data are examined, a slight decrease in all of the anthropometric index values is observed. This illustrates a decrease in the proportion of children malnourished after the multiple imputation application compared to the results based on observed values. The reason for this overall decrease can be related with the bias associated with the observed data, according to the results of the study. In other words, it can be stated that the missing portion of the child population under age five in TDHS-2003 are better nourished compared to the observed children, according to the results.

The approximate fractions of missing information in the multiple imputation analyses are substantially smaller than the percentages of missing data in the

observed data set. This implies the gains in precision from the imputation model, which makes use of information about missing data available in observed data is substantial, in the study (Khare et.al., 1993).

There are also some limitations encountered during the study concerned either with the multiple imputation procedure or variables selected, as well as some cautions for possible similar studies, which should be mentioned here. First of all, the type of variables to be imputed is important in the imputation process. The reason is that the multiple imputation procedure assumes multivariate normality, and for categorical variables such as gender, this can be a problematic issue. In this context, weight and height measurements were appropriate selections for this study because of two reasons; first one is the type of the variables are continuous, and the second one is that they have adequate amount of efficient predictors, as the small fractions of missing information confirm. It was an important advantage that the imputation models explained about 90 percent of the variation in both of the dependent variables, for the reliability of the models.

Imputing categorical variables has also some shortcomings when used as auxiliary variables as well. Although IVEware imputation procedure imputes these variables within the limitations of its own categories; during analyses stage the structure of the variables are distorted, because of taking the average of categories. In these situations, rounding the variable to the closest category is recommended in the literature. In this study, “sufficiency of antenatal care” and “size of child at birth” variables are rounded after the analyses stage.

The structure of the data is also a very important point in determining the imputation model. As an instance, TDHS-2003 is a hierarchical[10] data set in nature and there are

[10] In a hierarchical data model, data are organized in a relation based structure, ever-married women and children data sets in TDHS-2003 is an example for hierarchical data, such that a woman can have many children but a child can have only one mother.

many variables in TDHS-2003 that are structurally dependent to each other, which are related with the skip patterns of the questionnaires.

On the other hand, as mentioned shortly in Chapter 4, estimations regarding the other imputed variables in the study are inconvenient. The first reason for that is auxiliary variables are included in the data file for prediction purposes only. If they were to predict, probably many other different variables would be included to the imputation model. The second reason is the hierarchical structure of the data which is also mentioned above. The unit of data, which may be children under age five, women or live births of ever-married women, is an important feature that should be considered in the analyses. In line with this conclusion, these imputed auxiliary variables are generally not included in the final data sets for public usage, according to the literature.

Multiple imputation is a newly raising area for Turkey and there are many issues that could be addressed in future research. One important and fertile direction is the application of multiple imputation to other variables in the TDHS data sets, including continuous and categorical ones. Moreover, imputation of other waves of TDHS data can also be considered for comparisons and following important trends.

As well as that missing data is an ongoing problem in DHS data sets, not only for Turkey but also for all other implementing countries. Performing and development of more diversified and robust imputation techniques for DHS is another direction, which could also serve for more unbiased results and accurate comparisons between countries.

REFERENCES

Acuna, E. and Rodriguez, C. (2004), "The Treatment of Missing Values and Its Effect in the Classifier Accuracy", in D. Banks, L. House, F.R. McMorris, P. Arabie, W. Gaul (Eds). Classification, Clustering and Data Mining Applications, Springer-Verlag Berlin-Heidelberg.

Akadlı Ergöçmen, B., Coşkun, Y. and Eker, L. (2005), "Türkiye'de Doğum Öncesi Bakım ve Doğum Hizmetlerinden Yararlanma", In *2003 Türkiye Nüfus ve Sağlık Araştırması İleri Analiz Raporu*, Hacettepe Üniversitesi Nüfus Etütleri Enstitüsü, Ankara.

Allison, P. D. (2001), *Missing Data*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA, Sage.

Bailar, J.C. III and Bailar, B.A. (1978), "Comparison Of Two Procedures For Imputing Missing Survey Values", *Proc. Section on Survey Research Methods*, pp. 462-467.

Bal, C. and Özdamar, K. (2004), "Solving the Missing Value Problem by Use of Simulated Data Sets", *Osmangazi Üniversitesi Tıp Fakültesi Dergisi,* 2004: 26(2):67-76, Eskişehir, Turkey.

Buuren, S.V. and Oudshoorn K. (1999), "Flexible multivariate imputation by MICE", TNO Report.

Charmarbagwala, R., Ranger, N., Waddington, H. and White, H. (2004), *The Determinants of Child Health and Nutrition: A Meta-Analysis,* Department of Economics, University of Maryland and Operations Evaluation Department, World Bank, Washington, DC.

Centers for Disease Control and Prevention (1999), ANTHRO: Software for Calculating Pediatric Anthropometry, Version 1.02. (available at: http://www.cdc.gov/nccdphp/dnpa/anthro.htm).

Cohen, M.P (1996), "A New Approach to Imputation", 1996 *Proceedings of the Section on Survey Research Methods, American Statistical Association*, p. 293-298.

Cohen, M.P (2003), "Imputation Allowing Standard Variance Formulas", In *Journal of Data Science 1(2003)*, 275-292.

Collins, L.M., Schafer, J.L. and Kam, C.M (2001), "A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures", *Psyhological Methods*, Vol. 6, No: 4, 330-351.

Cox, B.G. and Folsom, R.E. (1978), "An Emprical Investigation of Alternative Item Nonresponse Adjustments", Proceedings of the Section on Survey Research Methods, Americal Statistical Association, pp. 219-223.

Cogill, B. (2003), Anthropometric Indicators Measurement Guide, Food and Nutrition Technical Assistance Project, Washington DC.

Croft, T. (1991), "DHS Data Editing and Imputation", *in Proceedings of the Demographic and Health Surveys World Conference*, Washington, D.C., 1991, Vol. II, 1337-1356, Columbia, Maryland.

Durrant, G. B. (2005), "Imputation Methods for Handling Item-Nonresponse in Social Sciences: A Methodological Review", NCRM Working Paper Series, 1-42.

Enders, C. K. (2006), "Analyzing Structural Equation Models with Missing Data", In G.R. Hancock & R.O. Mueller (Eds.), A Second Course in Structural Equation Modelling, Information Age: Greenwich, CT.

Ezzati-Rice, T., Khare, M., Rubin, D., Little, R. and Schafer, J. (1993), "A Comparison of Imputation Techniques in the Third National Health and Nutrition Examination Survey", *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Fay, R.E. (1991), "A design-based perspective on missing data variance", in Proceedings of the 1991 Annual Research Conference, U.S. Bureau of the Census.

Fay, R.E. (1992), "When are inferences from multiple imputation valid?", in Proceedings of the Survey Research Methods Section, American Statistical Association.

GSS (1996), "Report on the Task Force on Imputation", Government Statistical Service Methodology Series No. 3, UK.

Hu, M. and Salvucci, S. (2001), "A Study of Imputation Algorithms", National Center for Education Statistics, Working Paper No. 2001-17.

HUIPS (1994), *1993 Turkey Demographic and Health Survey*, Ministry of Health General Directorate of Mother and Child Health and Family Planning, Hacettepe University Institute of Population Studies and Macro International Inc., Ankara, Turkey.

HUIPS (1999), *1998 Turkey Demographic and Health Survey*, Ministry of Health General Directorate of Mother and Child Health and Family Planning, Hacettepe University Institute of Population Studies and Macro International Inc., Ankara, Turkey.

HUIPS (2004), *2003 Turkey Demographic and Health Survey*, Hacettepe University Institute of Population Studies, Ministry of Health General Directorate of Mother and

Child Health and Family Planning, State Planning Organization and European Union, Ankara, Turkey.

Kalton, G and Kasprzyk, D. (1986), "The Treatment of Mising Survey Data", In: Survey Methodology, June 1986, Vol. 12, No.1, Statistics Canada.

Khare, M., Little, R.J.A., Rubin, B. and Schafer, J.L. (1993), "Multiple Imputation of NHANES III", *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 1:297-302, Alexandria, VA: American Statistical Association.

Koç, İ. (2004), Appendix D: Quality of the Data, In *2003 Turkey Demographic and Health Survey*, Hacettepe University Institute of Population Studies, Ministry of Health General Directorate of Mother and Child Health and Family Planning, State Planning Organization and European Union, Ankara, Turkey, pp. 211-218.

Little, R. J. A. and Rubin, D. B. (1986), *Statistical Analysis with Missing Data*, John Wiley & Sons Inc., USA.

Little, R.J.A. and Rubin, D.B. (1987), "Nonresponse in Sample Surveys", in Statistical Analysis with Messy Data, John Wiley and Sons, pp.50-75.

Little, R.J.A. and Schenker, N. (1995), "Missing Data", in *Handbook of Statistical Modeling in the Social and Behavioral Sciences*, G. Arminger, C.C. Clogg, and M.E. Sobel, Eds, New York: Plenum, pp. 39-75.

Macro International (2004), "Description of the Demographic and Health Surveys Individual Recode Data File", Measure DHS+, USA.

Macro International (2007), Guidelines for the MEASURE DHS Phase II Main Survey Report, Calverton, Maryland, USA.

Meng, X.L. (1994), "Multiple-imputation inferences with uncongenial sources of input", Statistical Science.

Oğuzlar, A. (2001), "Alan Araştırmalarinda Kayıp Değer Problemi ve Çözüm Önerileri", Ulusal Ekonometri ve İstatistik Sempozyumu, Çukurova Üniversitesi İİBF Ekonometri Bölümü, Adana, Turkey.

Peng, C.Y., Liou, S.M. and Ehman, L.H. (2006), "Advances in Missing Data Methods and Implications for Social Studies Research", Indiana University-Bloomington.

Pigott, T.D. (2001), "A Review of Methods for Missing Data", *Educational Research and Evaluation*, Vol. 7, No. 4, pp. 353-383.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J. and Solenberger, P. (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models", *Survey Methodology*, Vol. 27, No. 1.

Raghunathan, T.E., Van Hoewyk, J. and Solenberger, P. (2002), *IVEware: Imputation and Variance Estimation Software User Guide*, Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan.

Rässler S., (2004), "Data Fusion: Identification Problems, Validity, and Multiple Imputation", Institute for Employment Research of the Federal Employment Services, Competence, Centre Empirical Methods.

Rubin, D.B. (1988), "An Overview of Multiple Imputation", *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 79-84.

Rutstein, S.O. and Johnson, K. (2004), "The DHS Wealth Index", DHS Comparative Reports No. 6, Calverton, Maryland, ORC Macro.

Rutstein, S.O. and Rojas, G. (2003), *Guide to DHS Statistics*, Demographic and Health Surveys, Calverton, Maryland, ORC Macro.

Schafer, J. L. (1999), "Multiple Imputation: A Primer", Statistical Methods in Medical Research, 8: 3-15.

Schafer, J.L., Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A. and Rubin, D.B. (1996), "The NHANES III Multiple Imputation Project", *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 1:28-37, Alexandria, VA: American Statistical Association.

Schafer, J.L. and Graham, J.W. (2002), "Missing Data: Our View of the State of Art", *Psychological Methods 2002*, Vol. 7, No: 2, 147-177.

Schenker, N., Raghunathan, T.E., Chiu, P.L, Makuc, D.M., Zhang, G and Cohen, A.J. (2006), "Multiple Imputation of Missing Income Data in the National Health Interview Survey", *Journal of the American Statistical Association*, Vol. 101, No. 475.

Schunk, D. (2007), "An Iterative Multiple Imputation Procedure for Dealing with Item Nonresponse in the German SAVE Survey", Mimeo, University of Mannheim.

Sinharay, S., Stern, H.S. and Russel, D. (2001), "The Use of Multiple Imputation fort he Analysis of Missing Data", Psychological Methods, Vol. 6, No. 4, 317-329.

Statistical Commission and Economic Commission for Europe (1999), "Model Explicit Item Imputation for Demographic Surveys and Censuses", Working Paper No. 30: 2-4, Rome, Italy.

Stasavage, D. (2005), "Democracy and Primary School Attendance Aggregate and Individual Level Evidence from Africa", Afro Barometer, Working Paper No. 54.

Tanguma, J. (2000), "A Review of the Literature on Missing Data", Paper presented at the Annual Meeting of the Mid-South Educational Research Association, 28th, Bowling Green, KY, November 15-17, 2000.

Taylor, J.M.G., Cooper, K.L., Wei, J.T., Sarma, R.V., Raghunathan, T.E. and Heeringa, S.G. (2002), "Use of Multiple Imputation to Correct for Nonresponse Bias in a Survey of Urologic Symptoms among African-American Men", *American Journal of Epidemiology*, 156, 774-782.

Türkyılmaz, A.S. (2003), *Estimation of Selected Demographic and Health Indicators for Provinces of Turkey from Census and Survey Data by Using Small Area Estimation Techniques*, Department of Technical Demography, Hacettepe University Institute of Population Studies, Unpublished Ph.D. Dissertation, May 2003, Ankara, Turkey.

Türkyılmaz, A.S., Hancıoğlu, A. and Koç, İ. (2004), Appendix B: Survey Design, In *2003 Turkey Demographic and Health Survey*, Hacettepe University Institute of Population Studies, Ministry of Health General Directorate of Mother and Child Health and Family Planning, State Planning Organization and European Union, Ankara, Turkey, pp. 167-185.

UN (2005), *Living Arrangement of Older Persons Around the World*, United Nations Department of Economic and Social Affairs, Population Division, New York.

Wayman, J.C. (2003), "Multiple Imputation For Missing Data: What Is It And How Can I Use It?", Paper presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago, IL.

Yansaneh, I.S., Wallace, L.S. and Marker, A.D. (1998), "Imputation Methods for Large Complex Datasets: An Application to the NEHIS", Proceedings of the section on Survey Research Methods, American Statistical Association, pp. 314-319.

Yuan, K.H. and Bentler, P.M. (2000), "Three Likelihood-Based Methods for Mean and Covariance Structure Analysis with Non-Normal Missing Data", in Sociological Methodology 2000:165-200. Washington, D.C. American Sociological.

APPENDIX A

Syntax for Multiple Imputation in IVEware

```
/* Multiply impute missing values in the study data file * /

libname thesisfile ‘c:\thesisfile’;

%IMPUTE (SETUP=NEW,NAME=syntax,DIR=c:\thesisfile);

DATAIN     thesisfile.datafile;

DATAOUT    thesisfile.dataout ALL;

DEFAULT    transfer;

CATEGORICAL sex v024 v025 v133 v190 enough sizechil;

CONTINUOUS  montfark reswght reshght breastfe wgtzero hgtzero
wgtoth hgtoth v001 pairpsu v005;

COUNT      v136;

RESTRICT wgtzero (months=0) hgtzero (months=0) wgtoth
(months>0) hgtoth (months>0);

BOUNDS     reswght (>=35,<=141) reshght (>=115,<=181) breastfe
(>=0,<=v012) enough (>=0,<3) sizechil (>0,<6) wgtzero
(>=3,<=5.2) wgtoth (>=2,<=34) hgtzero (>=49,<=59.3) hgtoth
(>=49,<=129);

ITERATIONS  4;

MULTIPLES   20;

SEED       1902;

RUN;
```

APPENDIX B

IVEware Output for 20 Multiple Imputations

Definitions of variables

Variable Name	**Variable Description**
v001	Cluster number
pairpsu	strata
v005	Sample weight
wgtzero	weight in kg for children aged “0”
wgtoth	weight in kg for other children
hgtzero	height in kg for children aged “0”
hgtoth	height in kg for other children
sex	Sex of child
montfark	Age in months
v012	Current age - respondent
v024	Region
v025	Type of place of residence
v133	Education in single years
v136	Number of household members
v190	Wealth index
reswght	Respondent's weight
reshght	Respondent's height
breastfe	Breastfeeding
enough	Sufficiency of antenatal care
sizechil	Size of child at birth

Imputation 1

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.020114	0.0475241
Std Dev	0.535413	0.314904	0.465711

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	21.6182	30.2
Mean	12.9246	10.99	12.8003
Std Dev	3.70987	6.00128	3.92563

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.212349	0.547745
Std Dev	2.64744	3.3003	5.32181

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	113.947	122
Mean	86.5483	76.39	85.7118
Std Dev	13.6829	31.4905	16.1569

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	35.4165	35.4165
Maximum	140	107.842	140
Mean	64.9246	62.841	64.8879
Std Dev	12.6048	13.5106	12.6226

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	144.246	115.8
Maximum	179.9	170.478	179.9
Mean	156.493	156.071	156.486
Std Dev	5.64606	6.15864	5.65487

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	1.83912	0
Maximum	56.5	31.884	56.5
Mean	11.3926	18.6038	11.419
Std Dev	8.39435	8.97692	8.40675

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	11	27.50	1144	26.17
1	1519	35.06	8	20.00	1527	34.93
2	1680	38.78	21	52.50	1701	38.91
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	1	7.14	65	1.49
2	636	14.59	5	35.71	641	14.66
3	2362	54.20	6	42.86	2368	54.16
4	743	17.05	1	7.14	744	17.02
5	553	12.69	1	7.14	554	12.67
Total	4358	100.00	14	100.00	4372	100.00

Imputation 2

Variable	Observed	Imputed	Double counted
V001	4372	0	0

PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.0200173	0.047428
Std Dev	0.535413	0.313274	0.464624

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	23.1393	30.2
Mean	12.9246	11.0835	12.8063
Std Dev	3.70987	5.97761	3.92061

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.2099	0.545312
Std Dev	2.64744	3.26085	5.29775

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	126.235	126.235
Mean	86.5483	76.8107	85.7465
Std Dev	13.6829	31.5781	16.1514

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	35.8013	35.8013
Maximum	140	91.7974	140
Mean	64.9246	66.2009	64.947
Std Dev	12.6048	11.0984	12.5798

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	145.941	115.8
Maximum	179.9	169.775	179.9
Mean	156.493	157.652	156.513
Std Dev	5.64606	5.37938	5.64298

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	6.2779	0
Maximum	56.5	28.4088	56.5

Mean	11.3926	16.0827	11.4097
Std Dev	8.39435	7.13768	8.39418

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	9	22.50	1142	26.12
1	1519	35.06	7	17.50	1526	34.90
2	1680	38.78	24	60.00	1704	38.98
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	1	7.14	65	1.49
2	636	14.59	1	7.14	637	14.57
3	2362	54.20	10	71.43	2372	54.25
4	743	17.05	1	7.14	744	17.02
5	553	12.69	1	7.14	554	12.67
Total	4358	100.00	14	100.00	4372	100.00

Imputation 3

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.0200375	0.0474481
Std Dev	0.535413	0.313571	0.464822

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	21.3837	30.2
Mean	12.9246	11.1133	12.8082
Std Dev	3.70987	5.98274	3.92027

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.210411	0.54582
Std Dev	2.64744	3.26902	5.30272

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372

Minimum	49.1	0	0
Maximum	122	120.508	122
Mean	86.5483	77.3131	85.7878
Std Dev	13.6829	31.6364	16.1385

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	39.9845	37.3
Maximum	140	97.9999	140
Mean	64.9246	63.3387	64.8966
Std Dev	12.6048	12.3732	12.6011

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	144.703	115.8
Maximum	179.9	171.049	179.9
Mean	156.493	156.028	156.485
Std Dev	5.64606	5.51291	5.64348

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	4.20351	0
Maximum	56.5	21.5885	56.5
Mean	11.3926	13.2007	11.3992
Std Dev	8.39435	6.12909	8.38737

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	10	25.00	1143	26.14
1	1519	35.06	10	25.00	1529	34.97
2	1680	38.78	20	50.00	1700	38.88
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	1	7.14	65	1.49
2	636	14.59	5	35.71	641	14.66
3	2362	54.20	6	42.86	2368	54.16
4	743	17.05	0	0.00	743	16.99
5	553	12.69	2	14.29	555	12.69
Total	4358	100.00	14	100.00	4372	100.00

Imputation 4

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.0201215	0.0475315
Std Dev	0.535413	0.314966	0.465753

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	22.322	30.2
Mean	12.9246	11.0922	12.8068
Std Dev	3.70987	5.92274	3.91502

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.212537	0.547933
Std Dev	2.64744	3.30327	5.32363

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	118.116	122
Mean	86.5483	76.858	85.7504
Std Dev	13.6829	31.4144	16.123

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	37.3584	37.3
Maximum	140	88.8016	140
Mean	64.9246	62.2055	64.8767
Std Dev	12.6048	12.2283	12.6019

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	139.793	115.8
Maximum	179.9	168.97	179.9
Mean	156.493	156.507	156.494
Std Dev	5.64606	5.40894	5.64143

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	2.68375	0
Maximum	56.5	22.2011	56.5
Mean	11.3926	12.7047	11.3974
Std Dev	8.39435	5.15736	8.38479

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	9	22.50	1142	26.12
1	1519	35.06	8	20.00	1527	34.93
2	1680	38.78	23	57.50	1703	38.95
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	1	7.14	65	1.49
2	636	14.59	5	35.71	641	14.66
3	2362	54.20	5	35.71	2367	54.14
4	743	17.05	3	21.43	746	17.06
5	553	12.69	0	0.00	553	12.65
Total	4358	100.00	14	100.00	4372	100.00

Imputation 5

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.0200719	0.0474823
Std Dev	0.535413	0.314182	0.465229

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	21.4147	30.2
Mean	12.9246	11.2508	12.817
Std Dev	3.70987	6.0297	3.92121

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.211283	0.546686
Std Dev	2.64744	3.28313	5.31131

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	116.04	122
Mean	86.5483	77.0282	85.7644
Std Dev	13.6829	31.4953	16.1283

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	39.8317	37.3
Maximum	140	95.2782	140
Mean	64.9246	67.1726	64.9642
Std Dev	12.6048	12.1579	12.5992

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	141.317	115.8
Maximum	179.9	178.164	179.9
Mean	156.493	157.532	156.511
Std Dev	5.64606	6.39621	5.66075

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	4.22336	0
Maximum	56.5	23.3184	56.5
Mean	11.3926	12.557	11.3968
Std Dev	8.39435	6.26988	8.38732

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	9	22.50	1142	26.12
1	1519	35.06	8	20.00	1527	34.93
2	1680	38.78	23	57.50	1703	38.95
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	2	14.29	66	1.51
2	636	14.59	2	14.29	638	14.59
3	2362	54.20	7	50.00	2369	54.19
4	743	17.05	3	21.43	746	17.06
5	553	12.69	0	0.00	553	12.65
Total	4358	100.00	14	100.00	4372	100.00

Imputation 6

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.0200027	0.0474134
Std Dev	0.535413	0.313121	0.464522

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	21.372	30.2
Mean	12.9246	10.9977	12.8008
Std Dev	3.70987	6.00678	3.92594

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.209528	0.544943
Std Dev	2.64744	3.25451	5.29389

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	117.173	122
Mean	86.5483	76.6631	85.7343
Std Dev	13.6829	31.5142	16.148

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	40.1448	37.3
Maximum	140	93.7476	140
Mean	64.9246	66.1898	64.9468
Std Dev	12.6048	11.3813	12.5842

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	147.025	115.8
Maximum	179.9	168.818	179.9
Mean	156.493	157.345	156.508
Std Dev	5.64606	5.23912	5.63978

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	1.11895	0
Maximum	56.5	28.4799	56.5
Mean	11.3926	15.2349	11.4066
Std Dev	8.39435	7.57734	8.39393

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	13	32.50	1146	26.21
1	1519	35.06	8	20.00	1527	34.93
2	1680	38.78	19	47.50	1699	38.86
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	1	7.14	65	1.49
2	636	14.59	1	7.14	637	14.57
3	2362	54.20	8	57.14	2370	54.21
4	743	17.05	4	28.57	747	17.09
5	553	12.69	0	0.00	553	12.65
Total	4358	100.00	14	100.00	4372	100.00

Imputation 7

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0

SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.0199853	0.0473962
Std Dev	0.535413	0.313159	0.464549

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	20.9597	30.2
Mean	12.9246	11.0562	12.8045
Std Dev	3.70987	6.0634	3.92981

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.20909	0.544507
Std Dev	2.64744	3.24826	5.29011

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	116.227	122
Mean	86.5483	76.9014	85.7539
Std Dev	13.6829	31.6867	16.1648

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	35.4202	35.4202
Maximum	140	98.7292	140
Mean	64.9246	65.1949	64.9293
Std Dev	12.6048	15.0533	12.65

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	144.562	115.8
Maximum	179.9	168.774	179.9
Mean	156.493	156.994	156.502
Std Dev	5.64606	5.17514	5.63804

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	2.11794	0
Maximum	56.5	20.2691	56.5
Mean	11.3926	12.9198	11.3982
Std Dev	8.39435	6.23361	8.38744

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	13	32.50	1146	26.21
1	1519	35.06	5	12.50	1524	34.86
2	1680	38.78	22	55.00	1702	38.93
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	3	21.43	67	1.53
2	636	14.59	4	28.57	640	14.64
3	2362	54.20	5	35.71	2367	54.14
4	743	17.05	2	14.29	745	17.04
5	553	12.69	0	0.00	553	12.65

Total	4358 100.00	14 100.00	4372 100.00

Imputation 8

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.0200364	0.047447
Std Dev	0.535413	0.313783	0.464964

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	24.0642	30.2
Mean	12.9246	11.1371	12.8097
Std Dev	3.70987	5.96017	3.91741

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.210384	0.545793
Std Dev	2.64744	3.26843	5.30235

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	113.46	122
Mean	86.5483	76.962	85.7589
Std Dev	13.6829	31.5991	16.148

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	41.7494	37.3
Maximum	140	87.7575	140
Mean	64.9246	61.994	64.873
Std Dev	12.6048	10.2028	12.5714

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	145.313	115.8
Maximum	179.9	170.015	179.9
Mean	156.493	156.749	156.498
Std Dev	5.64606	5.08698	5.63638

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	1.18488	0
Maximum	56.5	25.6666	56.5
Mean	11.3926	17.2179	11.4139
Std Dev	8.39435	6.52609	8.39507

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	13	32.50	1146	26.21
1	1519	35.06	9	22.50	1528	34.95
2	1680	38.78	18	45.00	1698	38.84
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	2	14.29	66	1.51
2	636	14.59	2	14.29	638	14.59
3	2362	54.20	7	50.00	2369	54.19
4	743	17.05	0	0.00	743	16.99
5	553	12.69	3	21.43	556	12.72
Total	4358	100.00	14	100.00	4372	100.00

Imputation 9

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.0201035	0.0475137
Std Dev	0.535413	0.314627	0.465527

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	21.7293	30.2
Mean	12.9246	11.085	12.8064
Std Dev	3.70987	5.95431	3.91829

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.212083	0.547481

Std Dev	2.64744	3.29512	5.31863

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	114.224	122
Mean	86.5483	76.8206	85.7473
Std Dev	13.6829	31.4588	16.1318

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	39.9418	37.3
Maximum	140	93.1815	140
Mean	64.9246	64.2727	64.9131
Std Dev	12.6048	13.0187	12.6109

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	146.163	115.8
Maximum	179.9	171.443	179.9
Mean	156.493	156.072	156.486
Std Dev	5.64606	5.59979	5.64489

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	0.0268498	0
Maximum	56.5	25.0997	56.5
Mean	11.3926	12.04	11.3949
Std Dev	8.39435	7.784	8.39146

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	11	27.50	1144	26.17
1	1519	35.06	9	22.50	1528	34.95
2	1680	38.78	20	50.00	1700	38.88
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	1	7.14	65	1.49
2	636	14.59	3	21.43	639	14.62
3	2362	54.20	8	57.14	2370	54.21
4	743	17.05	1	7.14	744	17.02
5	553	12.69	1	7.14	554	12.67
Total	4358	100.00	14	100.00	4372	100.00

Imputation 10

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0

ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.0201671	0.0475769
Std Dev	0.535413	0.315381	0.466029

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	23.3553	30.2
Mean	12.9246	11.1407	12.81
Std Dev	3.70987	6.0391	3.92505

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.213694	0.549082
Std Dev	2.64744	3.32255	5.33547

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	119.445	122
Mean	86.5483	76.8592	85.7505
Std Dev	13.6829	31.6333	16.1581

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	37.1527	37.1527
Maximum	140	97.8054	140
Mean	64.9246	65.4187	64.9333
Std Dev	12.6048	11.8055	12.59

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	140.252	115.8
Maximum	179.9	167.336	179.9
Mean	156.493	156.397	156.492
Std Dev	5.64606	5.30515	5.63975

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	1.29408	0
Maximum	56.5	23.9833	56.5
Mean	11.3926	14.2514	11.403
Std Dev	8.39435	6.8396	8.39032

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	9	22.50	1142	26.12
1	1519	35.06	11	27.50	1530	35.00
2	1680	38.78	20	50.00	1700	38.88
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	1	7.14	65	1.49
2	636	14.59	4	28.57	640	14.64
3	2362	54.20	6	42.86	2368	54.16
4	743	17.05	3	21.43	746	17.06

5	553	12.69	0	0.00	553	12.65
Total	4358	100.00	14	100.00	4372	100.00

Imputation 11

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.0200349	0.0474455
Std Dev	0.535413	0.313583	0.46483

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	21.7867	30.2
Mean	12.9246	10.966	12.7987
Std Dev	3.70987	5.95445	3.92177

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.210344	0.545754
Std Dev	2.64744	3.26846	5.30238

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	115.17	122
Mean	86.5483	76.514	85.722
Std Dev	13.6829	31.3254	16.1247

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	38.1138	37.3
Maximum	140	100.049	140
Mean	64.9246	62.315	64.8786
Std Dev	12.6048	13.7025	12.6279

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	142.824	115.8
Maximum	179.9	166.711	179.9

```
    Mean           156.493          156.205          156.488
 Std Dev           5.64606          5.74761           5.6473

Variable BREASTFE
                 Observed          Imputed         Combined
  Number             4356               16             4372
 Minimum                0           4.1751                0
 Maximum             56.5          31.4529             56.5
    Mean          11.3926          15.7824          11.4086
 Std Dev          8.39435          7.83823          8.39573

Variable ENOUGH
                 Observed          Imputed         Combined
    Code       Freq    Per       Freq    Per       Freq    Per
       0       1133  26.15         15  37.50       1148  26.26
       1       1519  35.06          7  17.50       1526  34.90
       2       1680  38.78         18  45.00       1698  38.84
   Total       4332 100.00         40 100.00       4372 100.00

Variable SIZECHIL
                 Observed          Imputed         Combined
    Code       Freq    Per       Freq    Per       Freq    Per
       1         64   1.47          2  14.29         66   1.51
       2        636  14.59          1   7.14        637  14.57
       3       2362  54.20          8  57.14       2370  54.21
       4        743  17.05          1   7.14        744  17.02
       5        553  12.69          2  14.29        555  12.69
   Total       4358 100.00         14 100.00       4372 100.00
```

Imputation 12

```
Variable          Observed          Imputed    Double counted
V001                  4372                0                 0
PAIRPSU               4372                0                 0
V005                  4372                0                 0
wgtzero                 28             4344                 0
wgtoth                4091              281                 0
hgtzero                 28             4344                 0
hgtoth                4012              360                 0
SEX                   4372                0                 0
MONTHS                4372                0                 0
MONTFARK              4372                0                 0
V012                  4372                0                 0
V024                  4372                0                 0
V025                  4372                0                 0
V133                  4372                0                 0
V136                  4372                0                 0
V190                  4372                0                 0
RESWGHT               4295               77                 0
RESHGHT               4296               76                 0
BREASTFE              4356               16                 0
ENOUGH                4332               40                 0
SIZECHIL              4358               14                 0

Variable wgtzero
                 Observed          Imputed         Combined
  Number               28             4344             4372
 Minimum                3                0                0
 Maximum              5.2              5.2              5.2
    Mean              4.3        0.0200831        0.0474934
 Std Dev         0.535413         0.314308         0.465313

Variable wgtoth
                 Observed          Imputed         Combined
  Number             4091              281             4372
 Minimum              2.2                0                0
 Maximum             30.2          20.7697             30.2
    Mean          12.9246           10.946          12.7974
 Std Dev          3.70987          5.93947          3.92092

Variable hgtzero
                 Observed          Imputed         Combined
  Number               28             4344             4372
 Minimum             49.4                0                0
```

Maximum	59.3	59.3	59.3
Mean	52.5821	0.211567	0.546968
Std Dev	2.64744	3.28773	5.31412

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	114.06	122
Mean	86.5483	76.9411	85.7572
Std Dev	13.6829	31.6028	16.1495

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	38.0351	37.3
Maximum	140	113.198	140
Mean	64.9246	65.0467	64.9267
Std Dev	12.6048	12.8695	12.608

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	138.436	115.8
Maximum	179.9	168.157	179.9
Mean	156.493	156.679	156.497
Std Dev	5.64606	6.10075	5.65357

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	7.30542	0
Maximum	56.5	35.9411	56.5
Mean	11.3926	16.6185	11.4117
Std Dev	8.39435	8.42099	8.39941

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	7	17.50	1140	26.08
1	1519	35.06	11	27.50	1530	35.00
2	1680	38.78	22	55.00	1702	38.93
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	1	7.14	65	1.49
2	636	14.59	1	7.14	637	14.57
3	2362	54.20	7	50.00	2369	54.19
4	743	17.05	5	35.71	748	17.11
5	553	12.69	0	0.00	553	12.65
Total	4358	100.00	14	100.00	4372	100.00

Imputation 13

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0

RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.0199374	0.0473486
Std Dev	0.535413	0.312486	0.464101

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	20.9121	30.2
Mean	12.9246	11.1159	12.8084
Std Dev	3.70987	6.08559	3.93033

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.207876	0.543302
Std Dev	2.64744	3.22699	5.27723

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	114.458	122
Mean	86.5483	76.8412	85.749
Std Dev	13.6829	31.6988	16.1694

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	39.2668	37.3
Maximum	140	96.1237	140
Mean	64.9246	66.1952	64.9469
Std Dev	12.6048	12.1259	12.5963

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	144.279	115.8
Maximum	179.9	172.326	179.9
Mean	156.493	157.163	156.505
Std Dev	5.64606	6.27918	5.65755

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	1.30789	0
Maximum	56.5	22.421	56.5
Mean	11.3926	12.637	11.3971
Std Dev	8.39435	5.83194	8.38627

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	13	32.50	1146	26.21
1	1519	35.06	5	12.50	1524	34.86
2	1680	38.78	22	55.00	1702	38.93
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	2	14.29	66	1.51
2	636	14.59	6	42.86	642	14.68

3	2362	54.20	5	35.71	2367	54.14
4	743	17.05	0	0.00	743	16.99
5	553	12.69	1	7.14	554	12.67
Total	4358	100.00	14	100.00	4372	100.00

Imputation 14

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.0199847	0.0473956
Std Dev	0.535413	0.312874	0.464358

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	22.5098	30.2
Mean	12.9246	11.1163	12.8084
Std Dev	3.70987	5.96868	3.91882

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.209073	0.54449
Std Dev	2.64744	3.24657	5.28908

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	115.303	122
Mean	86.5483	76.7927	85.745
Std Dev	13.6829	31.5941	16.1548

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	39.4965	37.3
Maximum	140	94.3103	140
Mean	64.9246	64.9066	64.9242
Std Dev	12.6048	11.8445	12.5905

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372

Minimum	115.8	139.678	115.8
Maximum	179.9	170.137	179.9
Mean	156.493	157.188	156.505
Std Dev	5.64606	5.9946	5.6523

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	3.09855	0
Maximum	56.5	20.7864	56.5
Mean	11.3926	12.1401	11.3953
Std Dev	8.39435	5.32749	8.38491

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	11	27.50	1144	26.17
1	1519	35.06	5	12.50	1524	34.86
2	1680	38.78	24	60.00	1704	38.98
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	1	7.14	65	1.49
2	636	14.59	2	14.29	638	14.59
3	2362	54.20	4	28.57	2366	54.12
4	743	17.05	5	35.71	748	17.11
5	553	12.69	2	14.29	555	12.69
Total	4358	100.00	14	100.00	4372	100.00

Imputation 15

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.0200524	0.0474629
Std Dev	0.535413	0.313798	0.464973

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	21.3578	30.2
Mean	12.9246	11.1766	12.8123
Std Dev	3.70987	5.98136	3.91841

Variable hgtzero

	Observed	Imputed	Combined

Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.210788	0.546195
Std Dev	2.64744	3.27486	5.30627

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	115.442	122
Mean	86.5483	77.1763	85.7766
Std Dev	13.6829	31.6286	16.1432

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	41.449	37.3
Maximum	140	91.3979	140
Mean	64.9246	66.1531	64.9462
Std Dev	12.6048	12.1793	12.5971

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	146.289	115.8
Maximum	179.9	177.142	179.9
Mean	156.493	157.673	156.514
Std Dev	5.64606	5.69853	5.64842

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	0.944275	0
Maximum	56.5	22.6873	56.5
Mean	11.3926	12.1394	11.3953
Std Dev	8.39435	7.39077	8.39027

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	13	32.50	1146	26.21
1	1519	35.06	8	20.00	1527	34.93
2	1680	38.78	19	47.50	1699	38.86
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	1	7.14	65	1.49
2	636	14.59	1	7.14	637	14.57
3	2362	54.20	10	71.43	2372	54.25
4	743	17.05	0	0.00	743	16.99
5	553	12.69	2	14.29	555	12.69
Total	4358	100.00	14	100.00	4372	100.00

Imputation 16

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0

V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.0201025	0.0475127
Std Dev	0.535413	0.314769	0.465621

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	21.5167	30.2
Mean	12.9246	11.0333	12.8031
Std Dev	3.70987	6.01033	3.92524

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.212057	0.547456
Std Dev	2.64744	3.29575	5.31902

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	119.807	122
Mean	86.5483	76.7363	85.7403
Std Dev	13.6829	31.7307	16.1794

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	41.2698	37.3
Maximum	140	98.9823	140
Mean	64.9246	67.2745	64.966
Std Dev	12.6048	11.4435	12.5878

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	143.268	115.8
Maximum	179.9	168.466	179.9
Mean	156.493	157.705	156.514
Std Dev	5.64606	5.57643	5.64645

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	0.319744	0
Maximum	56.5	30.4117	56.5
Mean	11.3926	16.3666	11.4108
Std Dev	8.39435	8.44348	8.39893

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	12	30.00	1145	26.19
1	1519	35.06	6	15.00	1525	34.88
2	1680	38.78	22	55.00	1702	38.93
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per

1	64	1.47	1	7.14	65	1.49
2	636	14.59	1	7.14	637	14.57
3	2362	54.20	9	64.29	2371	54.23
4	743	17.05	3	21.43	746	17.06
5	553	12.69	0	0.00	553	12.65
Total	4358	100.00	14	100.00	4372	100.00

Imputation 17

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.0199693	0.0473803
Std Dev	0.535413	0.31292	0.464389

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	20.784	30.2
Mean	12.9246	11.0626	12.8049
Std Dev	3.70987	5.83868	3.90777

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.208683	0.544103
Std Dev	2.64744	3.24006	5.28513

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	116.605	122
Mean	86.5483	76.8399	85.7489
Std Dev	13.6829	31.3327	16.1108

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	35.8783	35.8783
Maximum	140	100.932	140
Mean	64.9246	64.4479	64.9162
Std Dev	12.6048	13.1207	12.6126

Variable RESHGHT

	Observed	Imputed	Combined

Number	4296	76	4372
Minimum	115.8	147.442	115.8
Maximum	179.9	167.173	179.9
Mean	156.493	156.745	156.498
Std Dev	5.64606	4.5942	5.62911

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	1.8671	0
Maximum	56.5	27.9403	56.5
Mean	11.3926	16.0964	11.4098
Std Dev	8.39435	8.34353	8.39802

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	11	27.50	1144	26.17
1	1519	35.06	8	20.00	1527	34.93
2	1680	38.78	21	52.50	1701	38.91
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	1	7.14	65	1.49
2	636	14.59	1	7.14	637	14.57
3	2362	54.20	5	35.71	2367	54.14
4	743	17.05	5	35.71	748	17.11
5	553	12.69	2	14.29	555	12.69
Total	4358	100.00	14	100.00	4372	100.00

Imputation 18

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.0200928	0.047503
Std Dev	0.535413	0.314459	0.465414

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	21.8982	30.2
Mean	12.9246	11.1324	12.8094
Std Dev	3.70987	5.97365	3.91886

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.211812	0.547212
Std Dev	2.64744	3.29244	5.31701

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	116.925	122
Mean	86.5483	77.1034	85.7706
Std Dev	13.6829	31.5218	16.1292

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	41.1947	37.3
Maximum	140	93.2849	140
Mean	64.9246	62.7797	64.8868
Std Dev	12.6048	11.5558	12.589

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	142.049	115.8
Maximum	179.9	169.833	179.9
Mean	156.493	156.278	156.49
Std Dev	5.64606	5.91812	5.65026

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	1.43955	0
Maximum	56.5	25.3807	56.5
Mean	11.3926	14.9162	11.4055
Std Dev	8.39435	8.58672	8.39676

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	9	22.50	1142	26.12
1	1519	35.06	9	22.50	1528	34.95
2	1680	38.78	22	55.00	1702	38.93
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	1	7.14	65	1.49
2	636	14.59	1	7.14	637	14.57
3	2362	54.20	6	42.86	2368	54.16
4	743	17.05	3	21.43	746	17.06
5	553	12.69	3	21.43	556	12.72
Total	4358	100.00	14	100.00	4372	100.00

Imputation 19

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0

V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.0201336	0.0475436
Std Dev	0.535413	0.315043	0.465804

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	21.861	30.2
Mean	12.9246	11.1547	12.8109
Std Dev	3.70987	6.00286	3.9211

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.212845	0.548238
Std Dev	2.64744	3.30832	5.32673

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	116.711	122
Mean	86.5483	77.1229	85.7722
Std Dev	13.6829	31.615	16.1433

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	38.1998	37.3
Maximum	140	91.6121	140
Mean	64.9246	65.1535	64.9286
Std Dev	12.6048	12.1304	12.5953

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	146.056	115.8
Maximum	179.9	171.381	179.9
Mean	156.493	157.944	156.519
Std Dev	5.64606	5.74976	5.65039

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	3.34347	0
Maximum	56.5	25.9414	56.5
Mean	11.3926	12.648	11.3972
Std Dev	8.39435	6.51983	8.38802

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	11	27.50	1144	26.17
1	1519	35.06	7	17.50	1526	34.90
2	1680	38.78	22	55.00	1702	38.93
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	0	0.00	64	1.46
2	636	14.59	2	14.29	638	14.59
3	2362	54.20	8	57.14	2370	54.21
4	743	17.05	1	7.14	744	17.02
5	553	12.69	3	21.43	556	12.72
Total	4358	100.00	14	100.00	4372	100.00

Imputation 20

Variable	Observed	Imputed	Double counted
V001	4372	0	0
PAIRPSU	4372	0	0
V005	4372	0	0
wgtzero	28	4344	0
wgtoth	4091	281	0
hgtzero	28	4344	0
hgtoth	4012	360	0
SEX	4372	0	0
MONTHS	4372	0	0
MONTFARK	4372	0	0
V012	4372	0	0
V024	4372	0	0
V025	4372	0	0
V133	4372	0	0
V136	4372	0	0
V190	4372	0	0
RESWGHT	4295	77	0
RESHGHT	4296	76	0
BREASTFE	4356	16	0
ENOUGH	4332	40	0
SIZECHIL	4358	14	0

Variable wgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	3	0	0
Maximum	5.2	5.2	5.2
Mean	4.3	0.020041	0.0474515
Std Dev	0.535413	0.313633	0.464863

Variable wgtoth

	Observed	Imputed	Combined
Number	4091	281	4372
Minimum	2.2	0	0
Maximum	30.2	22.167	30.2
Mean	12.9246	10.9936	12.8005
Std Dev	3.70987	5.89726	3.91541

Variable hgtzero

	Observed	Imputed	Combined
Number	28	4344	4372
Minimum	49.4	0	0
Maximum	59.3	59.3	59.3
Mean	52.5821	0.210498	0.545907
Std Dev	2.64744	3.26944	5.30297

Variable hgtoth

	Observed	Imputed	Combined
Number	4012	360	4372
Minimum	49.1	0	0
Maximum	122	116.888	122
Mean	86.5483	76.8227	85.7475
Std Dev	13.6829	31.4662	16.1329

Variable RESWGHT

	Observed	Imputed	Combined
Number	4295	77	4372
Minimum	37.3	35.8712	35.8712
Maximum	140	104.195	140
Mean	64.9246	64.5137	64.9173
Std Dev	12.6048	13.1907	12.6139

Variable RESHGHT

	Observed	Imputed	Combined
Number	4296	76	4372
Minimum	115.8	148.182	115.8
Maximum	179.9	174.156	179.9
Mean	156.493	158.39	156.526
Std Dev	5.64606	5.41207	5.64692

Variable BREASTFE

	Observed	Imputed	Combined
Number	4356	16	4372
Minimum	0	0.10021	0
Maximum	56.5	19.0704	56.5
Mean	11.3926	9.30916	11.3849
Std Dev	8.39435	5.32587	8.38572

Variable ENOUGH

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
0	1133	26.15	13	32.50	1146	26.21
1	1519	35.06	8	20.00	1527	34.93
2	1680	38.78	19	47.50	1699	38.86
Total	4332	100.00	40	100.00	4372	100.00

Variable SIZECHIL

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	64	1.47	1	7.14	65	1.49
2	636	14.59	5	35.71	641	14.66
3	2362	54.20	3	21.43	2365	54.09
4	743	17.05	5	35.71	748	17.11
5	553	12.69	0	0.00	553	12.65
Total	4358	100.00	14	100.00	4372	100.00