Multiple Imputation: An Alternative Solution for Filling Data

Umamaheswari.D[1] Vijay Anand.R[2]
[1,2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]NGM College, Pollachi

Abstract— Missing values, common in epidemiologic studies, are a major issue in obtaining valid estimates. Missing data are often a problem in social science data. Imputation methods fill in the missing responses and lead, under certain conditions, to valid inference. Simulation studies have suggested that multiple imputation is an attractive method for imputing missing values, but it is relatively complex and requires specialized software. This article reviews several imputation methods used in the social sciences and discusses advantages and disadvantages of these methods in practice. Simpler imputation methods as well as more advanced methods, such as fractional and multiple imputation, are discussed. The paper introduces the implementation in either multiple or fractional imputation approaches. Software packages for using imputation methods in practice are reviewed highlighting newer developments. The paper discusses an example from the social sciences in detail, applying several imputation methods to a missing earnings variable. The objective is to illustrate how to choose between methods in a real data example. A simulation study evaluates various imputation methods, including predictive mean matching, fractional and multiple imputation. This article reviews various imputation methods used within the social sciences to compensate for item-nonresponse bias, and provides the best result in using Multiple Imputation.

Key words: Item-non response, imputation, simple imputation, multiple Imputations

I. INTRODUCTION

In sample surveys non-response is often a major problem. This is of particular concern in medical and social science data. Many researchers in the social sciences are often faced with nonresponse problems but may not be familiar with statistical analysis methods that address the missing data problem adequately. Variables of interest to social scientists, such as income variables, opinion and attitudes, beliefs and others might be regarded as sensitive questions and are therefore often prone to nonresponse. Dealing with nonresponse can be a difficult matter and it is important to apply adequate missing data methods to obtain valid inference.

II. MISSING DATA APPROACHES

By nonresponse it is meant that the required data are not obtained for all elements, which are selected for observation. Generally, a distinction is made between unit nonresponse, i.e. the failure of a selected sample member to respond, and item nonresponse where it is failed to obtain some required information from individual sample members. Unit nonresponse occurs if it is not possible to interview certain sample members or if sample members did not want to take part in the survey. Item nonresponse on the other hand occurs if the interviewer fails to ask a question, does not record the answer or the sample member refuses to answer a question or does not know the answer. There are several ways of dealing with nonresponse problems. For unit nonresponses normally weighting methods are applied. To compensate for item nonresponse a range of missing data methods exist, such as available case method, imputation methods, weighting methods and model-based procedures such as maximum likelihood estimation. The focus of this paper is on imputation methods to compensate for item-nonresponse. Nonresponse occurring in longitudinal studies, such as attrition or drop-out, will not be considered.

III. IMPUTATION TECHNIQUES

Imputation theory is constantly developing and thus requires consistent attention to new information regarding the subject. There have been many theories embraced by scientists to account for missing data but the majority of them introduce large amounts of bias. A few of the well-known attempts to deal with missing data include: hot deck and cold deck imputation; listwise and pairwise deletion; mean imputation; regression imputation; last observation carried forward; stochastic imputation; and multiple imputation.

IV. CASE DELETION

By far, the most common means of dealing with missing data is listwise deletion, which is when all cases with a missing value are deleted. If the data are missing completely at random, then listwise deletion does not add any bias, but it does decrease the power of the analysis by decreasing the effective sample size. If the cases are not missing completely at random, then listwise deletion will introduce bias because the sub-sample of cases represented by the missing data are not representative of the original.

Pairwise deletion involves deleting a case when it is missing a variable required for a particular analysis, but including that case in analyses for which all required variables are present. When pairwise deletion is used, the total N for analysis will not be consistent across parameter estimations. Because of the incomplete N values at some points in time, while still maintaining complete case comparison for other parameters, pairwise deletion can introduce impossible mathematical situations such as correlations that are over 100%.

V. IMPUTATION METHODS

Researchers modeling medical data often encounter the problem of missingness regarding one or more of the variablesunder investigation. The most common approach is todelete those observations with missing values, leading to acomplete participant analysis. This approach not only wastesdata and reduces power but also produces biased estimates. An alternative is to use one ofthe many methods available for imputing the missing values.

A. Single Imputation:

A once-common method of imputation was hot-deck imputation where a missing value was imputed from a randomly selected similar record. One form of hot-deck imputation is called "last observation carried forward", which involves sorting a dataset according to any of a number of variables, thus creating an ordered dataset. The technique then finds the first missing value and uses the cell value immediately prior to the data that are missing to impute the missing value. The process is repeated for the next cell with a missing value until all missing values have been imputed. In the common scenario in which the cases are repeated measurements of a variable for a person or other entity, this represents the belief that if a measurement is missing, the best guess is that it hasn't changed from the last time it was measured.

Cold-deck imputation, by contrast, selects donors from another dataset. Due to advances in computer power, more sophisticated methods of imputation have generally superseded the original random and sorted hot deck imputation techniques.

Another imputation technique involves replacing any missing value with the mean of that variable for all other cases, which has the benefit of not changing the sample mean for that variable. However, mean imputation attenuates any correlations involving the variable(s) that are imputed. This is because, in cases with imputation, there is guaranteed to be no relationship between the imputed variable and any other measured variables. Thus, mean imputation has some attractive properties for univariate analysis but becomes problematic for multivariate analysis.

Regression imputation has the opposite problem of mean imputation. A regression model is estimated to predict observed values of a variable based on other variables, and that model is then used to impute values in cases where that variable is missing. In other words, available information for complete and incomplete cases is used to predict whether a value on a specific variable is missing or not. Fitted values from the regression model are then used to impute the missing values. The problem is that the imputed data do not have an error term included in their estimation, thus the estimates fit perfectly along the regression line without any residual variance. This causes relationships to be over identified and suggest greater precision in the imputed values than is warranted. The regression model predicts the most likely value of missing data but does not supply uncertainty about that value.

Stochastic regression was a fairly successful attempt to correct the lack of an error term in regression imputation by adding the average regression variance to the regression imputations to introduce error. Stochastic regression shows much less bias than the above mentioned techniques, but it still missed one thing - if data are imputed then intuitively one would think that more noise should be introduced to the problem than simple residual variance.[3]

Although single imputation has been widely used, one shortcoming is it does not reflect the full uncertainty created by missing data. This problem is the motivation for "multiple imputation" as a method to give a full representation of the uncertainty that arises when data that were expected from an experimental situation are not observed.

B. Multiple Imputation:

In order to deal with the problem of increased noise due to imputation, Rubin (1987) developed a method for averaging the outcomes across multiple imputed data sets to account for this. The way this works is that imputation processes similar to stochastic regression are run on the same data set multiple times and the imputed data sets are saved for later analysis. Each imputed data set is analyzed separately and the results are averaged except for the standard error term (SE). The SE is constructed by the within variance of each data set as well as the variance between imputed items on each data set. These two variances are added together and the square root of them determines the SE, thus the noise due to imputation as well as the residual variance are introduced to the regression model.

Multiple imputation involves drawing values of the parameters from a posterior distribution. The posterior distribution reflects the noise associated with the uncertainty surrounding the parameters of the distribution that generates the data. Therefore the multiple imputations simulate both the process generating the data and the uncertainty associated with the parameters of the probability distribution of the data. More traditional methods like hot-deck imputation and Maximum-likelihood-based imputation fail to give a complete simulation of the uncertainty associated with missing data.

Of these, the method of multiple imputation (MI) isattractive since theoretical and simulation studies haveshown that it yields estimates with good statistical properties, such as efficiency and validity, when a correct model is specified for the imputation. However, MI is not wellunderstood in the medical community and requires advancedsoftware. Consequently, alternative, simpler methods of imputationare more commonly adopted at present. Multiple imputation is now a well-established technique for analyzing data sets where some unitshave incomplete observations. Provided the imputation model is correct, the resulting estimatesare consistent.

Almost all substantive data sets collected for social research contain missing data. Observationscan be missing by design (Little and Rubin, 1987, p. 9) or because, for one reason or another, theintended observations were not made. Motivated by work with social surveys, such problems ledRubin to develop multiple imputation, which is most fully described in Rubin (1987). There is nowa vast literature on the use of multiple imputation, and related data augmentation techniques forhandling missing data.

To our knowledge, systematic comparisons of methods ofimputation using real-life meta-data are lacking. In thisarticle, we compare the naive, complete participant methodand several imputation methods, including different implementationsof MI, for dealing with missing values.They provide an opportunity toapply different imputation methods to diverse epidemiologicdata.The idea of multiple imputation is that instead of filling in missing values to create a single imputed dataset, several imputed data sets are created each of which contains different imputed values. The analysis of a statistical model is then done on each of the imputed data sets. The multiple analyses are then combined to yield a single set of results. The major advantage of multiple imputation over single imputation is that it produces standard errors that reflect the degree of uncertainty due to the imputation missing values.

In general, multiple imputation techniques require that missing observations are missing at random (MAR).There are two major approaches to creating multiply imputed datasets. The first one is based on the joint distribution of all the variables in the imputation model, including variables to be imputed and variables to be used only for the purpose of imputing other variables. In this approach, the joint distribution of all variables in the imputation model is assumed to be multivariate normal. The other approach is based on each conditional density of a variable given other variables.

VI. Conclusion

Our study indicates that there were similar results when imputation used for replacing missing values compare to no imputation. However, the power of the study increased with different effect sizes when we replaced missing values. There are several methods for replacing missing value in the analysis. Single imputation method is due to bias if the proportion of missing value is large. Multiple imputations (MI) are another alternative method to replace missing value. Biased estimates, which are more likely to result from using data imputation methods otherthan multiple implicate, severely threaten the reproducibilityof research. By using the most sophisticated methods available, the quality of data analysis and associated reports of that research will be enhanced. This can strengthen the profession's knowledge base for practice, research and theory development.

References

[1] Scheuren, F. (2005). Multiple imputation: How it began and continues. The American Statistician, 59,: 315-319.Return
[2] Allison, P. D. (2002). Missing data (Sage University PaperSeries on Quantitative Applications in the SocialSciences 07-136).ThousandOaks, CA: SagePublications.
[3] Streiner, D. L. (2002).The case of the missing data:Methods of dealing with dropouts and otherresearch vagaries. Canadian Journal of Psychiatry, 47,68-75.