Imputation technique: replacing missing values in longitudinal data

Gladius Jennifer. H*

Department of Community Medicine, Karpaga Vinayaga Institute of Medical Sciences and Research Centre, Kancheepuram, Tamilnadu, India

Received: 13 August 2016
Accepted: 08 September 2016

*Correspondence:
Dr. H. Gladius Jennifer,
E-mail: gladiusjennifer@gmail.com

Copyright: © the author(s), publisher and licensee Medip Academy. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

In longitudinal studies, many cases are found missing in the follow-up data. These missing cases may arise due to item non response or unit non response. A data set with missing observations is often completed by using imputed values. The unit non response is usually carried out by some weighting adjustment. The item non response is made by the method called imputation. Imputation is a technique to replace a missing/incomplete or strange value with a more or less artificial value. There are plenty of methods available to impute the missing values in a longitudinal data. Imputation is useful because they make the data set easier to analyze, ensure consistency between the results from different analyses and reduce non response bias from item non response. But it is not necessary that the imputed value reduces the bias of the data, sometimes they may lead more bias also. It depends on the imputation procedure which we choose and also the form of estimate. The aim of this article is to sensitize doctors and post-graduate medical students about this useful analytical technique.

Keywords: Imputation, Non responses, Single imputation, Regression imputation

INTRODUCTION

In longitudinal study many cases found missing in the follow up data. These missing cases may arise due to item non response or unit non response. The unit non response occurs when no information is collected from a sample unit. The item non response occurs when some but not all the required information is obtained from the sample unit. A data set with missing observation is often completed by using imputed values. The unit non response is usually carried out by some weighting adjustments. The item non response is made by the method called imputation.

Imputation is a technique to replace a missing value with a more or less artificial value. It is a process of substituting data for incomplete or inconsistent items in surveys/longitudinal studies. Imputation techniques have both advantages as well as disadvantages. The main advantages are reduce biases and avoids increased variance; disadvantages are analysis is complex, overstating precision of estimates and lack of consistency in compensate of missing values in different subsets of data. Imputation techniques broadly classified as three types: single imputation, multiple imputation and regression imputation.

SINGLE IMPUTATION

Single imputation is defined as each missing value in a data set being filled in with one value yielding one complete data set. The single imputation method has the following types of imputation techniques: mean overall imputation, mean within cell imputation, random imputation, and random within cell imputation.

Mean overall imputation

It imputes the overall mean of the respondents to all missing values. If the independent variable x has no missing value then the imputed value of y is calculated as
\[ Y_{m}^{*} = \bar{y}_r \] If \( x \) has missing value in \( x \) is replaced with 
\[ x_{m}^{*} = \bar{x}_r \] Then the imputed value for \( y \) is calculated as 
\[ Y_{m}^{*} = \bar{y}_r \] 

**Mean within cell imputation**

This assigns each sampled unit to one of the mutually exclusive and exhaustive imputation cells. Within each cell the observed cell mean is assigned to each non respondent in the cell. If the independent variable \( x \) has no missing value then the imputed value of \( y \) is calculated as 
\[ Y_{m}^{*} = \bar{y}_r \] If \( x \) has missing value in \( x \) is replaced with 
\[ x_{m}^{*} = \bar{x}_r \] Then the imputed value for \( y \) is calculated as 
\[ Y_{m}^{*} = \bar{y}_r \]

**Random imputation**

Consider sample size \( n \) with \( m= n-r \) missing values. A random sample of size \( m \) is taken with replacement from \( r \) observed values. The selected respondents act as donors and their values are randomly assigned to the respondents. In this method the \( m \) imputed \( y \) values will vary from imputation to imputation. To obtain required values one must first take an expectation over repeated imputations and conditional on all the observed data \(((x_{i},y_{i});i=1 \to r, \{x_{nj}=r+1 \to n\})\) and then take an expectation over the model. 

**Random within cell imputation**

It is generalization of random imputation method. That is random within cell imputation method is nothing but random imputation method applied independently within each of the imputation cells. If \( x \) has any missing value then it is replaced with the mean. Then randomly selected respondent values are used to impute the missing value.

In the above four single imputation methods, mean within cell imputation better than overall imputation since it reduces bias when estimating the missing value within stratum. Same way random within cell imputation is better than others because missing values are randomly replaced by given respondent values.

**MULTIPLE IMPUTATIONS**

Multiple imputations is a technique that replaces each missing datum with a set of \( m>1 \) values. The \( m \) version of the complete data are analyzed by standard complete data methods and the results are combined using simple rules to yield estimates of standard errors and \( p \) value. The types of multiple imputations are EM algorithm and maximum likelihood function.

**REGRESSION IMPUTATION**

Regression imputation is defined as replacing the missing values with predicted values from the estimated regression model. Regression based imputation are a natural extension of mean imputation. Instead of fitting only a constant term, a host variable, which predict non response are used. Rather than assuming that non response is completely random, regression based imputation assumes that non response is not at random but rather occurs solely along observed lines.

The types of regression imputations are simple linear regression prediction imputation, random regression imputation, two step method, regression based nearest neighbour hot deck ing method etc.

**Simple linear regression prediction imputation**

If suppose we have simple linear model \( Y \) on \( X = Y = \beta_0 + \beta_1 X + \varepsilon \) then the missing values replaced by the equation \( Y^* = \beta_0 + \beta_1 X \) when the independent variable has no missing values. If the independent value has the missing value it is calculated by \( \bar{X} = \bar{X} \). Then imputed value is calculated by using the formula \( Y^* = \beta_0 + \beta_1 \bar{X} \).

**Random regression imputation**

Random regression imputation imputes values directly from the estimated regression line. Regression residual errors can be added to the regression prediction to provide dispersion about the regression line. The method imputes missing value \( Y = Y^* = \bar{Y} + \varepsilon \) where \( \varepsilon = Y - \beta_0 + \beta_1 X \). if independent variable \( X \) is missing \( \bar{X} = \bar{X} \) the only difference between simple linear regression imputation and random regression imputation is \( \varepsilon \). Random is better than simple. SLR imputation under estimate the outcome values than random regression imputation, more or less similar to before imputation.

There are some other models also available in imputation; they are two step models, regression based nearest neighbour hot decking, sequential hot deck method, random hot deck method and conditional hot deck method.

This review will give a conclusion that, random imputation will be better technique in replacing missing values than by mean or median values of the variables. In single imputation random within cell imputation and for regression fit random regression imputation will be appropriate for longitudinal studies. There are various software available to calculate imputation such as SPSS, R code, SAS.

**CONCLUSION**

Imputation is useful because they make the data set easier to analyze ensure consistency between the results from different analyses and reduce non response bias from item non response. But it is not necessary that the imputed value reduces the bias of the data sometimes they may lead to more bias also. It depends on the imputation procedure which we choose and also the form of estimate. If the missing values not so many in the data then the imputation technique is useful to replace the missing values. Otherwise it leads to bias and inappropriate to the data.
Funding: No funding sources
Conflict of interest: None declared
Ethical approval: Not required

REFERENCES

1. Schafer JL. Analysis of incomplete multivariate data, EM and Data augmentation. London: Chapman and Hall; 1997: 37-83.
2. Rubin DB. Multiple Imputations for nonresponse in surveys. Newyork NY: John Willey & Sons; 1987.
3. Diggle PJ, Liang KY, Zeger SL. Missing values in longitudinal data, Analysis of Longitudinal Data. Newyork : Oxford Science Publications;1992: 208-221.
4. Dwyer JH, Feinleib M, Lippert P, Meister H. Statistical Models for longitudinal studies of health, Missing data in longitudinal studies. Newyork: Oxford science publications; 1992: 277 – 297.
5. Jinn JH. The effect of different imputation methods on analytical statistics of simple linear regression. Interstat; 2002.
6. Yim C. California Polytechnic State University, San Luis Obispo, Imputing Missing Data using SAS® http://support.sas.com/resources/papers/proceedings 15/3295-2015.pdf. Accessed on 10 January 2016.
7. IBM SPSS Software; http://www-03.ibm.com/software/products/en/spss-missing-values. Accesses on 20 February 2016.

Cite this article as: Jennifer GH. Imputation technique: replacing missing values in longitudinal data. Int J Community Med Public Health 2016;3: 2709-11.