EUROPEAN ACADEMIC RESEARCH
Vol. VI, Issue 8/ November 2018

Impact Factor: 3.4546 (UIF)
DRJI Value: 5.9 (B+)

ISSN 2286-4822
www.euacademic.org

A Comparative Study between Multiple Imputation Method and Regression Imputation Method of Estimation of Missing Data

MONTASIR A. A. MOHAMMED
College of Business and Economics, Qassim University
Kingdom of Saudi Arabia
ADIL M. Y. WANISS
College of Business and Economics, Qassim University
Kingdom of Saudi Arabia
KHALID R. K. GENAWI
College of Science, Sudan University of Science and Technology,
Republic of Sudan

Abstract

Statistical procedures for missing data have vastly improved, yet misconception and unsound practice still abound for missing data, and as with other statistical methods, missing data often create major problems when estimating of parameters. This paper discusses results of multiple imputation method and regression imputation method of estimation of missing data. These methods have statistical properties that are almost as good can be applied to a much wider array of models and estimation methods. Based on results of study, we concluded there is no a statistically significant difference between means of estimates of multiple imputation method and regression imputation method, and The best method was multiple imputation where the MAE were lower than that of regression imputation method.

Key words: missing data, mean absolute error MAE, multiple imputation method, regression imputation method.

Theoretical formulation

In this Section, we will introduce some basic concepts that will be used in the rest of the paper. These concepts include:

Definition of missing data

Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data [1]. Accordingly,

Some studies have focused on handling the missing data, problems caused by missing data, and the methods to avoid or minimize such in medical research [2,3].

Missing data mechanisms

Little and Rubin [4] outline three missing data mechanisms:

1. Missing completely at random (MCAR). If data are missing under this mechanism then it is as if random cells from the rectangular data set are not available such that the missing values bear no relation to the value of any of the variables.

2. Missing at random (MAR). Under this mechanism, missing values in the data set may depend on the value of other observed variables in the data set, but that conditional on those values the data are missing at random. The key is that the missing values do not depend on the values of unobserved variables.

3. Not missing at random (NMAR). Describes the case where missing values do depend on unobserved values**.**

Imputation methods of missing data

Imputation is where the missing data can be replaced with statistical estimates of the missing values. The goal of any imputation technique is to produce a complete data set that can then be analyzed using statistical methods for complete data.

Several methods exist for imputing missing values. These are described in more detail below.

Firstly: regression imputation method

A much more promising method is to use standard regression analysis to provide estimates of the missing data conditional on complete variables in the analysis. For example, for the simple case of univariate missingness in a single continuous variable Y, we fit a regression model to explain Y by the remaining p variables represented by the vector X using the complete cases (subscripted by i):

$$Y_i = \alpha + \sum_{k=1}^{p} \beta_k X_{ik} + \varepsilon_i \qquad (1)$$

Predicted values for the expected values of the missing cases of Y (subscripted by j) can be obtained from

$$\hat{Y}_j = \hat{\alpha} + \sum_{k=1}^{p} \hat{\beta}_k X_{jk} \qquad (2)$$

It should be emphasized that the equations above could be generalized to include models for non-continuous data such as binomial or count data.

Missing data are usually multivariate and it is possible to extend the procedure of regression based imputation from the univariate case to deal with multivariate missingness. For each missing value in the data set a model can be fitted for that variable employing the complete cases of all the other variables [5]. Where the number of variables with missing values is large, the number of models to be fitted will also be large, however, efficient computational methods (such as Little & Rubin's sweep operator) can be employed [4]. Alternatively, an iterative regression approach can be adopted [6] whereby missing values in a given variable are predicted from a regression of that variable on the complete cases of all other variables in the dataset. This process is repeated for all variables with missing values using complete cases of the other variables *including*

previously imputed values until a completed rectangular data set has been generated. The imputation of missing values for each variable is then re-estimated in turn using the complete set of data and the process continues until the imputed values stop changing

Advantages of regression imputation

The imputation retains a great deal of data over the listwise or pairwise deletion and avoids significantly altering the standard deviation or the shape of the distribution. However, as in a mean substitution, while a regression imputation substitutes a value that is predicted from other variables, no novel information is added, while the sample size has been increased and the standard error is reduced.[7]

Secondly: The multiple imputation method

It is important to recognize that when employing any imputation method we are estimating a missing value that is not observed. It is straightforward to see that in the case of unconditional mean imputation, the variance of the completed variable will be too low, since the imputed means do not contribute to the variance. However, the same is true with the other forms of imputation – if the expected value of the missing data point is imputed, although this is the 'best' prediction of the missing value (in the sense of mean squared error), there will be no allowance for the uncertainty associated with the imputation process. For example, if imputations are based on a regression equation, as in Equation (2) for the simple univariate missingness example, then there will be no variation between predicted values for observations with the same values for all of the other non-missing variables. Such 'deterministic' imputation approaches [6] will therefore underestimate the variance of any estimators in subsequent statistical analysis of the imputed data set. Therefore, imputed values of missing

data should include a random component to reflect the fact that imputed values are estimated (using so-called 'stochastic' imputation methods [6]) rather than treating the imputed values as if they are known with certainty.

For the regression example, two components to the uncertainty in the imputation process can be distinguished. The first component is the mean squared error from the regression which represents the between observation variability not explained by the regression model. Two approaches to including this error term are either: to select a value at random from a normal distribution with variance equal to the mean squared error from the regression; or to compute the residuals from the regression and to add one of these residuals at random to each of the imputed values from the regression. Of these two approaches, the second non-parametric bootstrap approach is probably preferred since it is straightforward to do and does not rely on the parametric assumption of normally distributed errors. The second component of uncertainty comes from the fact that the coefficients of the regression model are themselves estimated rather than known. The variance of the prediction error for each covariate pattern can be obtained from the variance–covariance matrix and, assuming multivariate normality, this component of uncertainty can also be incorporated into the stochastic imputation procedure.

Clearly, once missing values are imputed with a random component, then a complete data set will no longer be unique and the results of any analysis of will be dependent on the particular imputed values. The principle of multiple imputation uses this fact directly in order to allow estimation of variance in statistics of interest in an analysis that include representation of uncertainty in the true values of the missing information.

With multiple imputation, an incomplete data set will have the missing values imputed several (M) times, where the values to fill in are drawn from the predictive distribution of the

missing data, given the observed data. Each imputed data set is then separately analyzed with the desired methods for complete data. The variability in the statistic of interest across the alternative data sets then gives an explicit assessment of the increase in variance due to missing data. Thus this variance of each final parameter estimate is composed of two parts: the estimated variance within each imputed data set and the variance across the data sets.

Suppose that the statistic of interest in the analysis is given by y.

The steps in the multiple imputation procedure are then:

1. Generate M sets of imputed values for the missing data points, thus creating M completed data sets.
2. For each completed data set, carry out the standard complete data analysis, obtaining estimate $\hat{\theta}_i$ of interest and its estimated variance $\hat{a}r(\hat{\theta}_i)\ for\ i = 1 \ldots M$.
3. Combine the results from the different data sets. The multiple imputation estimate of θ is

$$\hat{\theta} = \frac{1}{M}\sum_{I=1}^{M}\hat{\theta}_i$$

(i.e. the mean across the imputed data sets) and multiple imputation estimate of variance is

$$v\hat{a}r(\hat{\theta}) = \frac{1}{M}\sum_{i=1}^{M} v\hat{a}r(\hat{\theta}_i) + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_{i=1}^{M}(\hat{\theta}_i - \hat{\theta})$$

The first term on the right hand side of this equation relates to the variance within the imputed data sets, whereas the term on the far right captures the uncertainty due to the variability in the imputed values, i.e. between the imputed data sets. The term $1+1/M$ is a bias correction factor.

The approximate reference distribution for interval estimates and significance tests is a t distribution with degrees of freedom $= (M-1)(1+r^{-1})^2$; [8] where r is the estimated

ratio of the between-imputation component of variance (numerator) to the within-imputation component of variance (denominator).

Rubin [9] shows that the relative efficiency of an estimate based on M complete data sets to one based on an infinite number of them is approximately $(1 + \gamma/M)^{-1}$ where γ is the rate of missing data. With 50% missing data, an estimate based on M ¼ 5 complete data sets has a standard deviation that is only about 5% wider than one based on infinite M. Unless rates of missing data are very high, there is little advantage to using more than five complete data sets [10].

Advantages of multiple imputation method:
It removes some of its limitations. Multiple imputation can be used with any kind of data and model with conventional software. When the data is MAR, multiple imputation can lead to consistent, asymptotically efficient, and asymptotically normal estimates[11]

Application
In this Section, we will introduce the applied side

RESULTS AND DISCUSSION

With respect to means and Std. deviation for estimates multiple imputation method and estimates regression imputation method

We calculated means and std. deviation, to know is there ostensibly difference between means. Table (1) shows that

Table (1): Comparison of Mean, Std. Deviation and Std. Error Mean results between estimates of multiple imputation method and estimates of regression imputation method

missing data	Multiple imputation		Regression imputation	
	Mean	Std. deviation	Mean	Std. deviation
10%	1002.48	7.56	1002.20	8.32
20%	1003.20	6.73	1002.82	11.44
30%	1001.81	7.94	1001.35	11.05

Source: The researcher from applied study, SPSS Package, 2018

This study revealed that the multiple imputation method has means, greater than means of regression imputation method. And the std. deviations of multiple imputation method lower than the std. deviations of regression imputation method.

We note from the table (1) there is ostensibly difference between means for estimates of multiple Imputation method and estimates of regression imputation method, and to know the statistical significance of differences, we used t-test.

Table (2) shows that
With respect to t-test
To test this hypothesis, we calculated values of t and p-value, table (2) shows that

Table (2): t-test, p-value and Mean Difference

missing data	t-test		
	t	p-value	mean difference
10%	0.075	0.941	0.268
20%	0.128	0.899	0.381
30%	0.188	0.852	0.467

Source: The researcher from applied study, SPSS Package, 2018

From the above table, it shows the p-values for t-test of the missing data of 10%, 20% and 30% are respectively (0.941), (0.899) and (0.852) which are greater than significant level (0.05). And, therefore, there is no a statistically significant difference between means.

With respect to Std. Error Mean and MAE

We calculated std. error mean and MAE depend on estimates of multiple imputation method and estimates of regression imputation method. Table (3) shows that

Table (3): Std. Error Mean and MAE

Missing data	Multiple imputation		Regression imputation	
	Std. Error Mean	MAE	Std. Error Mean	MAE
10%	2.39	6.76	2.63	8.70
20%	1.51	9.67	2.56	13.05
30%	1.45	8.37	2.02	10.08

Source: The researcher from applied study, SPSS Package, 2018

The results of this study revealed that the std. error mean calculated by estimated data of multiple imputation method was lower than the std. error mean calculated by estimated data of regression imputation method. These results were consistent with MAE of multiple imputation method which was also lower than MAE of regression imputation method. Hence, based on those results, we concluded to multiple Imputation method best than regression imputation method in estimation of missing data.

REFERENCES

1. Graham JW. 2009; Missing data analysis: making it work in the real world. Annu Rev Psychol. 60:549-576.
2. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. 2012;The prevention and treatment of missing data in clinical trials. N Engl J Med. 367:1355–1360
3. O'Neill RT, Temple R. 2012; The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it. Clin Pharmacol Ther. 91:550–554.

4. Little RJA, Rubin DB. Statistical Analysis with Missing Data. John Wiley & Sons, New York, 1987.
5. Buck SF. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. J Roy Stat Soc Series B 1960; 22(2): 302–306.
6. Brick JM, Kalton G. Handling missing data in survey research. Stat Meth Med Res 1996; 5: 215–238.
7. Hyun Kang. The prevention and handling of the missing data. Korean J Anesthesiol. 2013 May; 64(5): 402–406
8. Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable non-response. J Am Stat Assoc 1986; 81: 366–374.
9. Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley: New York, 1987.
10. Schafer JL. Multiple imputation: a primer. Stat.
11. Schafer, J. L. 1997. Analysis of Incomplete Multivariate Data, New York: Chapman and Hall.