INVESTIGATING THE EFFECTS OF IMPUTATION NUMBERS ON VARIANCE OF ESTIMATES

Nwakuya, M. T. PhD
Department of Mathematics / Statistics
University of Port Harcourt, P.M.B 5323
Port Harcourt, **NIGERIA**
&
Nwabueze Joy C. PhD
Department of Statistics, Michael Okpara
University of Agriculture Umudike
P.M.B. 7267, Abia State, **NIGERIA**

ABSTRACT

The number of imputations to use in a multiple imputation analysis is always in question, knowing that missing data is a major problem in most research works. Most software's have five number of imputations as their default setup for any datasets with any percentage of missing values. Some researchers recommend that percentage of missing data should equate the number of imputations. But this is argued, as high number of missing values will attract very high number of imputations, which will take more computing time. Multiple imputation method imputes multiple values into a single missing point generating multiple complete data sets. In this paper we compared the variances of multiple regression estimates gotten from complete data sets imputed using 6 different imputation numbers, namely 50, 40, 30, 20, 15 and 5.The sample sizes investigated are 20000, 8000 and 30, each having 30%, 20% and 10% missing values. This work was analyzed in R software. Each of the complete datasets was analyzed and results pooled to give a single inference. The variances of the estimates were compared to each other to determine if they were significantly different from each other based on the imputation number used to impute the missing values and the percentage of missing value. The paired comparison was done in SPSS and the analysis showed that the variances were not significantly different from each other irrespective of the number of imputation used. But when it was compared based on the percentage of missing values, the variances were found to be significantly different.

Keywords: Imputation number, Missingness, Comparison, Variance and Multiple Imputation.

INTRODUCTION

Researchers are often faced with missing values in their investigations, this could arise due to informants refusing or forgetting to answer survey questions, files getting lost or data not recorded properly. Given the expense of collecting data, waiting to develop full proof methods of gathering information seems unattainable. Many methods have been developed to tackle missing data; these include complete case analysis, available case analysis, single value imputation, multiple imputation analysis amongst others. In this paper we will consider the multiple imputation analysis, were data is missing in all variables in a non monotone form. Multiple Imputations involves imputing the missing points a number of times to get complete datasets which are analyzed individually and pooled together to get a single inference. Multiple Imputations acknowledge the uncertainty stemming from filling in missing values rather than observing them, Rubin (1987) and Schafer, (1997). In this paper, we tried to check if there is a significant difference between the variances of the estimates

ISSN 2056-5879

gotten from datasets imputed using different imputation numbers. We considered 6 different imputation numbers namely 50, 40, 30, 20, 15 and 5 numbers of imputations.

LITERATURE REVIEW

Ignorability

In most research works especially in biomedical research works, only the parameters of the distribution of repeated measure are of interest, while those related to missingness pattern are viewed as nuisance parameters. When inference about the measurement mechanism can be made without explicitly addressing the missingness mechanism then missingness is considered to be ignorable Geert M. and Geert V. (2005). Rubin, (1976) earlier simplified the definition by stating that missingness is considered ignorable if the missingness mechanism is independent of the observed given the missing. Rubin, (1978) classified ignorability into two, namely Missing completely at random (MCAR) and Missing at Random (MAR).

Imputation

An alternative way to obtain a data set on which complete data method can be used is to fill in rather than delete, Little and Rubin (1987). Filling in implies imputing, imputation can be classified into single and multiple imputations. In single imputation a value is substituted for every missing value in the data set and the resulting data set is analyzed as if it represents the true complete data. No units are excluded from the analysis, thus the original number of included units is maintained at all points. Single imputation omits possible differences between multiple imputations, single imputation will tend to underestimate the standard errors and thus overestimate the level of precision. Thus, single imputation gives the researcher more apparent power than the data justify. While Multiple Imputation replaces each missing value with a set of *m* plausible values, the imputed datasets are then analyzed using standard procedures for complete data and combining the results from these analyses to get a single inference. Multiple imputation is a principled missing data method that provides valid statistical inferences under Missing at Random condition, Rubin, (1978). Allison P, (2012) stated that there is need to have more than one imputed data set because only one imputed data set gives highly inefficient estimates.

Imputation numbers

Allison P. (2012) stated that over the last decade, multiple imputation has rapidly become one of the most widely used methods for handling missing data. He said however, one of the big uncertainties about the practice is how many imputed data sets are needed to get good results. Graham et al (2007) recommended 20 imputations for 10% to 30% missing values and 40 imputations for 50% missing values. Similar recommendations were proposed by Bodner (2008) and Royston et al (2011). The agreed that the number of imputations to use should depend on the percentage of missing values. The argument is that if the number of missing value is very high then too many imputations will be needed, increasing the imputation time. According to Carpenter and Kenward, (2013) and Va Buuren, (2012), in other to reduce the effect of simulation error we need to increase the number of imputations and this will also reduce the variance of the estimates. They recommended the number of imputations to be 50 or more.

ISSN 2056-5879

METHODOLOGY

Based on the recommendations of Bodner (2008) and Royston et al (2011), that the percentage of missing values should equate the number of imputations, we applying the shrinkage estimator proposed by Nwakuya and Nwabueze (2016), tried to compare the variances of regression estimates from the complete dataset gotten from 6 different imputations, to see if the number of imputations and the percentages of missing value affects the variances.

Procedure

Three different regression data sets of sample size 20000, 8000 and 30 with 30%, 20% and 10% missing values for all datasets were simulated in R software. The simulated regression data had 3 independent variables and each independent variable had missing values in a non-monotone pattern. Applying the shrinkage estimator proposed by Nwakuya and Nwabueze (2016), given by, where is the shrinkage parameter. Given that is the regression coefficient, is the independent variable, m no of imputations and. We obtained the following results.

RESULTS:

Table 4.1: Comparison of Imputation Variance across the sample sizes based on the Imputation numbers

Sample sizes	Percentage missing	Imputation number 50	Imputation number 40	Imputation number 30	Imputation number 20	Imputation number 15	Imputation number 5
	30%	20764614	21450124	21354953	24137531	25343514	15995333
20000	20%	13934849	13908126	14089326	13669698	14862953	13114601
	10%	9518349	9598227	9497460	9529869	9484578	10145260
	30%	30676.65	29646.41	29718.84	30899.52	30452.88	39985.64
8000	20%	26438.19	26295.48	26222.32	27741.83	28099.31	30717.62
	10%	24753.22	24989.61	25449.22	26253.05	26125.7	26496.67
	30%	11501.63	11641.09	11622.81	10643.81	12612.59	13162.31
30	20%	10739.21	10764.55	11071.13	10756.69	10404.26	12410.95
	10%	9651.271	9634.434	9663.463	9820.533	9766,661	9638.835

Table 4.2 Paired Sample test based on imputation numbers

		Paired Differences					T	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	imp50 - imp40	-81986.2670	228178.65806	76059.5527	-257379.9100	93407.37601	-1.078	8	.312
Pair 2	imp50 -	-80434.956	198296.86129	66098.9537	-232859.41	71989.50382	-1.21	8	.258

	imp30	9			76		7		
Pair 3	imp50 - imp20	-346849.029	1138215.3451	379405.115	-1221758.793	528060.7351	-.914	8	.387
Pair 4	imp50 - imp15	-1692647.62	3377176.7788	1125725.59	-4288575.491	903280.2536	-1.504	8	.171
Pair 5	imp50 - imp5	549329.572	1623888.5506	541296.14	-698901.6658	1797560.809	1.015	8	.340
Pair 6	imp40 - imp30	1551.31011	80640.87477	26880.2916	-60434.75345	63537.37367	.058	8	.955
Pair 7	imp40 - imp20	-264862.762	911887.66125	303962.554	-965801.6680	436076.1438	-.871	8	.409
Pair 8	imp40 - imp15	-1610661.35	3315392.1366	1105130.71	-4159097.344	937774.6405	-1.457	8	.183
Pair 9	imp40 - imp5	631315.839	1840406.4917	613468.831	-783345.8214	2045977.499	1.029	8	.334
Pair 10	imp30 - imp20	-266414.072	954010.98250	318003.661	-999731.8291	466903.6847	-.838	8	.426
Pair 11	imp30 - imp15	-1612212.66	3322522.0731	1107507.36	-4166129.209	941703.8847	-1.456	8	.184
Pair 12	imp30 - imp5	629764.529	1820918.7528	606972.918	-769917.5293	2029446.587	1.038	8	.330
Pair 13	imp20 - imp15	-1345798.59	3197179.9959	1065726.67	-3803368.687	1111771.508	-1.263	8	.242
Pair 14	imp20 - imp5	896178.600	2732995.2172	910998.406	-1204587.489	2996944.692	.984	8	.354
Pair 15	imp15 - imp5	2241977.19	4195878.2103	1398626.07	-983260.3107	5467214.692	1.603	8	.148

ISSN 2056-5879

Table 4.3: Comparison of Imputation Variance across Imputation numbers based on the percentages of missingness for n=20,000

Imputation numbers	30%	20%	10%
50	20764614	13934849	9518349
40	21450124	13908126	9598227
30	21354953	14089326	9497460
20	24137531	13669698	9529869
15	25343514	14862953	9484578
5	15995333	13114601	10145260

Table 4.4: Paired Sample test based on % of missingness for n=20,000

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	30% - 20%	7577752.67	2810660.0675	1147447.1676	4628145.8203	10527359.5131	6.604	5	.001
Pair 2	30%-10%	11878721.0	3465111.7697	1414625.9562	8242309.2127	15515132.7873	8.397	5	.000
Pair 3	20%-10%	4300968.33	782055.13109	319272.67031	3480251.8064	5121684.86025	13.471	5	.000

Table 4.5: Comparison of Imputation Variance across Imputation numbers based on the percentages of missingness for n=8,000

Imputation numbers	30%	20%	10%
50	30676.65	26438.19	10739.21
40	29646.41	26295.48	10764.55
30	29718.84	26222.32	10764.55
20	30899.52	27741.83	10756.69
15	30452.88	28099.31	10404.26
5	39985.64	30717.62	12410.95

Table 4.6: Paired Sample test based on % of missingness for n=8,000

		Paired Differences					t	Df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	30% - 20%	4310.86500	2502.96134	1021.82969	1684.16816	6937.56184	4.219	5	.008
Pair 2	30% - 10%	20923.2883	3305.49800	1349.46391	17454.3809	24392.1957	15.505	5	.000
Pair 3	20% - 10%	16612.4233	1226.27084	500.62297	15325.5310	17899.3156	33.184	5	.000

Table 4.7: Comparison of Imputation Variance across Imputation numbers based on the percentages of missingness for n=30

Imputation numbers	30%	20%	10%
50	9518349	24753.22	9651.271
40	9598227	24989.61	9634.434
30	9497460	25449.22	9663.463
20	9529869	26253.05	9820.533
15	9484578	26125.7	9766,661
5	10145260	26496.67	9638.835

Table 4.8: Paired Sample test based on % of missingness for n=30

		Paired Differences					T	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	30% - 20%	9603279.25500	255654.3932	104370.468	9334986.4233	9871572.08667	92.01	5	.000
Pair 2	30%-10%	7993112.24400	4061462.126	1658084.969	3730869.1393	12255355.3486	4.821	5	.005
Pair 3	20%-10%	-1610167.0110	3983050.799	1626073.679	-5790122.476	2569788.45370	-.990	5	.368

DISCUSSION

In this paper we tried to investigate the effects of imputation number and missing values on variance of estimates. Applying 6 different imputations on 3 different sample sizes with 3 different percentages of paired across all the number of imputations. This implies that the number of imputations does not affect the estimates. Furthermore comparing the variance across the imputation number based on the percentage of missingness for all the sample sizes, we discovered that the variance were significantly different from each other. This goes to confirm that missing values in a data set affects the estimates. From tables 4.3, 4.5 and 4.7 we observe that the variance were highest when missingness was 30% and lowest when missingness was 10% irrespective of the number of imputation used in imputing the missing values. We also noticed in table 4.6 that comparison between 20% and 10% missingness for sample size 30 was not significant, this we can attribute to the fact that the sample size was small.

CONCLUSION

We conclude based on the analysis that missing values affect estimates. The more missing values we have in a dataset the more the variance of the estimates, irrespective of the number of imputations used in the analysis. We also conclude that the number of imputations does not affect the estimates.

ISSN 2056-5879

REFERENCES

Allison P. (2012), "Why you Probably Need More Imputation Than you Think" www.statisticalhorizon.com. Accessed 13th May 2016.

Bodner, T E. (2008), "What Improves with Increased Missing Data Imputation?" *Structural Equation Modeling*: A Multidisciplinary Journal 15, 651-675.

Graham, John W., Allison P, Olchowski E. and Tamika D. G. (2007), "How many Imputations Are Really Needed?" Some Practical Clarifications of Multiple Imputation Theory", *Prevention Science* 8: 206-213.

Geert M.and Geert V. (2005), "Models for Discrete Longitudinal Data" Springer-Verlag, NY, 567-578.

Nwakuya M. T, and Nwabueze J. C. (2016), "Relative Efficiency of Estimates Based on Percentages of Missingness Using Three Imputation Numbers in Multiple Imputation Analysis",*European Journal of Physical and Agricultural Science*, vol 4, No 1.

Rubin D. B. (1978), "Multiple Imputation in sample surveys- a phenomenological Bayesain approach to nonresponse. In imputation and editing of Faulty or Missing Survey Data". Washington D C: US Department of Commerce.

Rubin D. B. (1976), "Inference and Missing Data", *Biometrika*, 63. 581-592.

Rubin D. B. (1987), "Multiple Imputation for Non-response in Surveys", JohnWiley and Sons, New York, 546-550.

Schafer J. L. (1997), "Analysis of Incomplete Multivariate Data", Chapman & Hall, London, pp 87-95

Royston P., White, Ian R and Wood M. A. (2011), "Multiple Imputation using Chained Equations": Issues and Guidance for Practice". *Statistics in Medicine* 30: 377-399.