RELATIVE EFFICIENCY OF ESTIMATES BASED ON PERCENTAGES OF MISSINGNESS USING THREE IMPUTATION NUMBERS IN MULTIPLE IMPUTATION ANALYSIS

Nwakuya, M. T. (Ph.D)
Department of Mathematics/Statistics
University of Port Harcourt P.M.B
5323, Port Harcourt Rivers State
NIGERIA

Nwabueze Joy C.
Department of Statistics Michael
Okpara University of Agriculture
Umudike P.M.B.7267, Abia State
NIGERIA

ABSTRACT

Most researchers have faced the problem of estimation when data points are missing. The mostly adopt easy to implement procedures without considering the efficiency of their estimates. In this paper we looked at the relative efficiency of estimates in Multiple Imputation analysis, based on percentages of missing data using 3 different imputation numbers; 7, 5 and 3 on four different simulated data sets with 50%, 45%, 25% and 10% missing values. The variance of each data set with different percentages of missing value for each imputation number was computed using a proposed method. This proposed method was seen to yield lower variances compared to an existing method. The program was written and implemented in R. The pooled variance of the estimates was also computed based on the percentages of missing values in the different data sets. The relative efficiency were computed and compared among the 3 different imputation numbers using the T-test for paired sample test in SPSS. From the results it was observed that when the missingness was 50% the estimates from data set gotten from imputation number 7 was most efficient when compared to estimates from data sets gotten from imputation numbers 5 and 3. When the missingness was 10% and 25% the estimates from data set gotten from imputation number 5 were found to be most efficient followed by estimates from data sets gotten from imputation number 7 and then 3. The relative efficiency for 40% missingness compared among the 3 imputation numbers showed that estimates from imputation number 3were most efficient.

Keywords: Multiple Imputation, Relative Efficiency, Imputation Variance, Missing Values and Shrinkage Parameter.

INTRODUCTION

Missing data is defined as data value that should have been recorded but for some reasons was not, Molenberg G, Verbeke G. (2005). Most researchers have faced the problem of missing quantitative data at some point in their work. Missing data is a potential source of bias in every analysis according to the European Agency for Evaluation of Medical Products (2001). Missing data leave us with the decision of how to analyse data when we do not have complete information from all informants. When information is missing in a sample, some researches employ any easy to administer method without checking the efficiency of their estimates. This paper considers the relative efficiency of estimates from data imputed using 3 different imputation numbers in a multiple imputation analysis. We will focus on these sets of data with different percentages of missing values. Multiple Imputation is a principled missing data method that provides valid statistical inferences under Missing at Random condition, Rubin (1978), Tanner and Wong (1987), Rubin and Schenker (1986) and Schafer's (1997). We applied a proposed Shrinkage estimator in this analysis that yielded lower variances compared to Ordinary least square estimates. In this paper the missing data pattern applied is the Multivariate non-monotone missing pattern; this is a situation where data points are missing randomly from more than one variable.

LITERATURE REVIEW

Missing data concept

There are three main missing data mechanism described by Rubin (1976) namely Missing Completely At Random (MCAR), this is when the probability of an observation being missing is independent of the responses; Missing At Random (MAR), this is said to be a condition in which the probability that data are missing depends only on the observed values, but not the missing values, after controlling for the observed and Missing Not At Random (MNAR), here the probability of a measurement being missing depends on unobserved data. Dong and Peng (2013), stated that there are three patterns of missing data, namely: univariate, monotone and non-monotone (arbitrary) missing patterns. Suppose there are m variables denoted as, $X_1, X_2, \dots, X_m$, a data set is said to have a univariate missing pattern if the missing data is from only one of the m variables and if in more than one variable, it is multivariate missing pattern. A data set is said to have a monotone missing data pattern, if the variables can be arranged in such a way that, when X_j is missing $X_{j+1}, X_{j+2}, \dots, X_m$ are also missing as well. Non-monotone missing data pattern occurs when more than one of the m variables has missing data points in a random manner. Many researchers use ad hoc methods such as complete case analysis, available case analysis (pairwise deletion), or single-value imputation. Though these methods are easily implemented, they require assumptions about the data that rarely hold in practice T.D. Pigott, (2001).

Multiple Imputation

According to Rubin (1987), Multiple Imputation analysis involves three stages namely: The missing values are filled in M times to generate M complete data sets; The M complete data sets are analyzed by using standard procedures;The results from the M analyses are combined into a single inference. According to Carpenter J. R. and Kenward M. G. (2013), also Va Burren (2012), in other to reduce the effect of the simulation error we need to increase M (number of imputations).

Estimators

Tony ke, (2012), gave an insight on measuring the goodness of an estimator. He said that intuitively an estimator is good, if it is close to the unknown parameter of interest or the estimator error is small. In the context of estimating regression coefficients Stein (1956) proposed ashrinkage estimator that dominates the ordinary least squares. Anchoring on Stein's discovery Ohtani (2009), compared a shrinkage estimator and OLS estimator for regression coefficient. . Lebanon G, (2006) stated that the relative efficiency of two unbiased estimators is the ratio of their variances. The quality of two estimators can be compared by looking at the ratio of their MSE. If two estimators are unbiased it is equivalent to the ratio of the variances which is defined as the relative efficiency, Lebanon, G. (2006).

METHODOLOGY

Our motivation stems from the use of high imputation numbers in other to reduce the effect of simulation error in multiple imputation analysis as proposed by Carpenter J. R. and Kenward M. G. (2013), and also from the regression coefficient estimator with a shrinkage parameter proposed by Ohtani K. (2009). We essentially restrict our data distribution to be normally distributed with multivariate non-monotone missingness.

Proposed method

This regression coefficient proposed by Ohtani K. (2006) is given by;

$$\beta_* = \left(1 - \frac{ae'e}{\beta' S\beta}\right)\beta, \qquad (1)$$

Where, $e = y - \boldsymbol{X}b;\ 0 \leq a \leq \frac{2(k-2)}{n-k+2}\ (n\ is\ the\ sample, k\ is\ number\ of\ parameters); s = X'X$

$\beta = (X'X^{-1})X'Y$ & $\left(1 - \frac{ae'e}{\beta' S\beta}\right)$ *is the shrinkage parameter where a is an abitrary number*

Our proposed shrinkage estimator is given by; $\hat{\beta}^*_{new} = \left(1 - \frac{(m-2)\tau}{\beta' X' X\beta}\right)\beta$

We introduced a parameter $\tau = \frac{e'e}{n-m}$ into equation (1)

$, where\ n\ is\ the\ sample\ size\ and\ m\ is\ the\ number\ of\ imputation, e = y - \boldsymbol{X}b\ and$ $\beta\ is\ the\ ordinary\ least\ square\ imputation\ estimate.$

Procedure

A program was written in R to implement this new approach. Four different data sets of sample size n = 30, 500,1000, 5000 &10000 were simulated with 10% 25%, 40% and 50% missing values. The missimg data points were imputed using imputation numbers 3, 5 and 7 for each sample size. The proposed estimator was applied in Multiple Imputation analysis to obtain the total imputation variances which were lower than the ones from ordinary least square estimates. We then applied the relative efficiency given by $\frac{var(\theta_1)}{Var(\theta_2)}$, Lebanon G.(2006)……………………..……..(2)

Where we have,

$\frac{var(\theta_1)}{Var(\theta_2)} = \frac{n+2}{3}\ for\ n > 1$, then θ_2 has a lower variance thus more efficient than θ_1.

The pooled variance is given by $s_p = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2+\cdots+(n_5-1)s_5^2}{n_1+n_2+n_3+n_4+n_5-k}$ …………(3)

Given that $n_1 = 10{,}000;\ n_2 = 5000;\ n_3 = 1000;\ n_4 = 500;\ n_5 = 30;$ k=5 (number of sample sizes) and s_i^2 are the individual variances. We used the T test for comparison of paired means in SPSS software to compare the variances gotten from estimates from data sets imputed using the three imputation numbers.

RESULTS

Table 3.1: Total imputation variances for each imputation number

TOTAL VARIANCES FROM THE PROPOSED METHOD			TOTAL VARIANCES FROM THE METHOD		
IMPUTATION NUMBER 7	IMPUTATION NUMBER 5	IMPUTATION NUMBER 3	IMPUTATION NUMBER 7	IMPUTATION NUMBER 5	IMPUTATION NUMBER 3
40190	43386.7	71266.93	40192.82	43389.95	71274.14
27242.11	23131.12	21023.61	27243.12	23131.57	21023.76
27054.68	24217.74	21881.38	27055.65	24218.32	21881.62
22490.76	22373.06	24115.34	22491.04	22373.32	24115.85
61293.45	68699.28	48008.12	61298.89	68699.28	48010.02
63499.24	55041.41	55699.91	63505.37	55045.45	55703.77
50019.25	50023.33	72463.67	50021.57	50025.71	72471.63
47941.48	45748.96	46393.13	47943.29	45750.1	46394.47
272058.4	300450.8	303103.9	272146.8	300578.2	303234.4
258556.4	290739.4	207539.1	258653.1	290889.3	207560.2
236626.8	251912	274405.7	236689.5	252001.3	274529.7
232796.9	231889.7	235831.4	232836.9	231928.5	235876.7
814832.6	697774.5	465811	815943.8	698587.1	465820.6
453141.4	476997.8	438336.6	453420	477322.4	438514.2
409747.6	425971.1	365321.6	409887.2	426139.8	365329.1
383069	378149.5	375767.8	3831012	378162.1	375775.4
57207.16	52034.29	99600.58	58134.3	52985.4	101515.5
12770.14	11665.19	19530.32	12779.86	11700.43	19565.53
3288.36	3645.03	5580.666	3333.9	3704.12	5701.69
1947.402	1991.24	1937.624	1947.3	1991.24	1937.59

Table 3.2: Comparison of the total imputation variances among the 3 imputation numbers

Paired Sample T test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	VarImp7 - VarImp3	16107.738	81918.309	18317.491	-22231.2110	54446.686	.879	19	.390
Pair 2	VarImp5 - VarImp3	15111.189	58902.509	13171.002	-12456.034	42678.411	1.147	19	.265
Pair 3	VarImp7 - VarImp5	996.54910	29744.597	6651.0941	-12924.3519	14917.449	.150	19	.882

Table 3.3: Pooled variances

Imputation Numbers	Pooled Variances for all percentages of missingness			
	50% missingness	40% mssingnes	25% missingness	10% mssingness
7	84,012.864	65,029.488	58,185.5107	53,755.9199
5	86,360.069	63,010.16	57,884.7281	52,818.1006
3	90,209.956	55,388.079	62,791.3112	54,233.491

Table 3.4: Relative Efficiency for 50% missingness

Imputation numbers		Relative Efficiency
Pair 1	VarImp7 & VarImp5	$\frac{Var\ (Imp\ 7)}{Var(Imp\ 5)}$ = 0.9728
Pair 2	VarImp7 & VarImp3	$\frac{Var\ (Imp\ 7)}{Var(Imp\ 3)}$ = 0.9313
Pair 3	VarImp5 & VarImp3	$\frac{Var\ (Imp\ 5)}{Var(Imp\ 3)}$ = .9573

Table 3.5 Relative Efficiency for 40% missingness

Imputation numbers		Relative Efficiency
Pair 1	VarImp7 & VarImp5	$\frac{Var\ (Imp\ 7)}{Var(Imp\ 5)}$ = 1.0321
Pair 2	VarImp7 & VarImp3	$\frac{Var\ (Imp\ 7)}{Var(Imp\ 3)}$ = 1.1741
Pair 3	VarImp5 & VarImp3	$\frac{Var\ (Imp\ 5)}{Var(Imp\ 3)}$ = 1.1376

Table 3.6: Relative Efficiency for 25% missingness

Imputation numbers		Relative Efficiency
Pair 1	VarImp7 & VarImp5	$\frac{Var\ (Imp\ 7)}{Var(Imp\ 5)}$ = 1.005
Pair 2	VarImp7 & VarImp3	$\frac{Var\ (Imp\ 7)}{Var(Imp\ 3)}$ =0.9267
Pair 3	VarImp5 & VarImp3	$\frac{Var\ (Imp\ 5)}{Var(Imp\ 3)}$ = 0.9219

Table 3.7: Relative Efficiency for 10% missingness

Imputation numbers		Relative Efficiency
Pair 1	VarImp7 & VarImp5	$\frac{Var\ (Imp\ 7)}{Var(Imp\ 5)}$ = 1.0177
Pair 2	VarImp7 & VarImp3	$\frac{Var\ (Imp\ 7)}{Var(Imp\ 3)}$ =0.9912
Pair 3	VarImp5 & VarImp3	$\frac{Var\ (Imp\ 5)}{Var(Imp\ 3)}$ = 0.9739

DISCUSSION

We begin with the imputation variances. Looking at table3.1, we observe that the new imputation variance from our proposed method is seen to be lower than that from the ordinary least square method. From the paired t-test in table 3.2, we discovered that there is no significant difference between the new total variances from all the three number of imputations. This goes to show that the reduction in the total variance was not due to increase in number of imputations but can be attributed to the improved method, irrespective of the number of imputations. From the relative efficiency results it was observed that when the missingness was 50% the estimates from data set gotten from imputation number 7 was most efficient when compared to estimates from data sets gotten from imputation numbers 5 and 3. When the missingness was 10% and 25% the estimates from data set gotten from imputation number 5 were found to be most efficient followed by estimates from data sets gotten from imputation number 7 and then 3. The relative efficiency for 40% missingness compared among the 3 imputation numbers showed that estimates from imputation number 3were most efficient.

CONCLUSIONS

In conclusion, generally our proposed method produced lower variances compared to the ordinary least square method and we observed that this reduction is not due to any increase in the number of imputations but it was based on the new approach. We found out that for large sample sizes with moderate missing values, imputation number 7 was most appropriate for achieving efficient estimates, while for low missing values imputation numbers 5 and 3 can be used.

REFERENCES

Carpenter J. R. and Kenward M. G. (2013), *Multiple Imputation and its Application*, John Wiley and Sons, Ltd. Publication, 37- 73.

Dong Y. and Peng C J. (2013), Principled Missing Data Methods for Researchers. *Springer Plus*, 2:22. http:www.springerplus.com/content/2/1/222.

European Agency for the Evaluation of medicinal products, 2001, *Evaluation of Medicines for Human Use*. www.ema.europa.eu/ema/pages/includes/documents/open_document.jsp?...

Lebanon G. (2006), *Relative Efficiency, Efficiency and the Fisher Information*, www.cc.gatech.edu/~lebanon/notes/efficiency.pdf

Molenberghs G and Verbeke G (2005) , *Models for Discrete Longitudinal Data,* Springer-Verlag, NY, 567-578.

Ohtani K (2009), *Comparison of some shrinkage estimators and OLS estimator for regression coefficients under the Pitman nearness criterion:* A Monte Carlo Study, Kobe University Economic Reviews, 55.

Pigott T. D. (2001), *A Review of Methods of Missing Data*, Educational Research & Evaluation. Taylor & Francis, 100-112,353-383.

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

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

Rubin D. B. and Schenker N. (1986), Multiple Imputation of Interval estimation from Simple Random samples with ignorable nonresponse, *Journal of the American Statistical Association*, pp 97-102.

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

Stein C. (1956), *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*, Proceeding of the third Berkeley Symposium on Mathematical Statistics and Probability ,1, Berkeley University of California press, vol1, 197-206.

Tanner and Wong (1987), "The Calculation of Posterior Distribution by Data Augmentation", *Journal of American Statistical Association*, 82, 528-550.

Tony ke (2012), *James Stein Estimator*, www.ieor.berkeley.edu/~kete/uploads/1/2/4/0/12408873/js.pdf