International Journal of Statistics and Applications 2018, 8(6): 305-308
DOI: 10.5923/j.statistics.20180806.03

Comparison of Shrinkage–Based Estimators in the Presence of Missing Data: A Multiple Imputation Analysis

M. T. Nwakuya[1,*], J. C. Nwabueze[2]

[1]Department of Mathematics/Statistics, University of Port Harcourt, Port Harcourt Rivers, Nigeria
[2]Department of Statistics, Michael Okpara University of Agriculture Umudike, Abia State, Nigeria

Abstract In this paper we examined the performance of the mean square error of the Ordinary Least Square (OLS) estimator, Minimum Mean Square Error (MMSE) estimator, N/N shrinkage Estimator (N/NSE) and a proposed Adjusted Minimum Mean Square Error (PAMMSE) estimator in a multiple imputation analysis when data points are missing in different data sets. The program for the proposed adjusted minimum mean square error was written and implemented in R. It is shown by numerical computations that the PAMMSE Estimator seem to be the best choice among OLS, MMSE, N/NSE and PAMMSE estimators in terms of their mean square errors when applied in multiple imputation analysis.

Keywords Mean square error, Shrinkage estimator, Imputation numbers, Multiple Imputation, Missingness

1. Introduction

Missing data is always a major concern in most data analysis. Awareness has grown of the need to go beyond complete case analysis of datasets with missing data points, following the work of Rubin [5] and [4]. Complete case analysis basically means deleting every missing points. This method leads to reduction of sample sizes which invariably reduces the degree of freedom. In other to avoid the short falls of complete case analysis and other missing data methods, the multiple imputation method was introduced by [5]. The basic idea of data analysis with multiple imputations is to create 'm' different copies of a data each of which has its missing value suitably imputed. These complete data sets are each analysed independently. The estimates of interest are averaged across the 'm' copies to give a single estimate. In most analysis the number of multiple imputations is not usually considered during parameter estimation. [2] Observed the need to incorporate the number of imputations in parameter estimation and investigated Ohtani's shrinkage estimator. Ohtani [3] proposed a shrinkage estimator for regression estimates, hinging on [7] stein's shrinkage estimator. This [7] Stein's shrinkage estimator was found to dominate Ordinary Least square estimator in terms of mean square error. [2] Proposed anextension of Ohtani's shrinkage estimator to multiple imputation analysis, their Shrinkage estimator also showed to have a lower mean square error than the Ordinary Least Square method. As one of the shrinkage estimators for regression coefficients, [8] Proposed the minimum mean square error (MMSE). [3] Derived the exact formula of the mean square error of the minimum mean square error (MMSE) estimator and showed that minimum mean square error (MMSE) dominates the Ordinary Least Square estimator in terms of mean square error. Further [3] proposed an adjusted minimum mean square error (AMMSE) estimator and showed that it has a lower mean square error than minimum mean square error (MMSE) estimators. In this work our focus is on the adjusted minimum mean square error estimator and how it can be extended to incorporate the number of multiple imputations.

Mean square error is arguably the most important criterion used to evaluate the performance of an estimator. It is calculated as t sum of the variance of the estimator and the squared bias of the estimator; this relationship is given by

$$MSE = E_{\hat{\theta}}\left(\left(\theta - \hat{\theta}\right)^2\right) = B_{\hat{\theta}}\left(\theta\right)^2 + V_{\hat{\theta}}\left(\theta\right) \qquad (1)$$

The smaller the mean square error the lesser the variability and the better the estimator; [1].The quality of an estimator can be assessed by computing and assessing its

* Corresponding author:
tobenwakuya@gmail.com (M. T. Nwakuya)
Published online at http://journal.sapub.org/statistics

mean square error. Though recently the Pitman's nearness criterion is gaining more ground in assessing the quality of an estimator.

In this work we tried to extend the adjusted minimum mean square error estimator to multiple imputation analysis, by incorporating the number of imputations into the formula and then comparing its results to that of the Ordinary least Square estimator and the other two shrinkage estimators namely the minimum mean square error estimator and the Nwakuya/Nwabueze Shrinkage estimator by [2]. The aim of this work is to compare the proposed estimator to the other different estimators based on their mean square errors. The objectives are to simulate normally distributed regression data sets with missing values, suitably input the missing values and compute the variances of the estimates based on the number of imputations. The program for this work was written and implemented in R. The paired comparison was done in SPSS at 0.05 level of significance.

2. Estimators

Given a regression model;

$$y = X\beta + \epsilon \tag{2}$$

$y_{nx1} \rightarrow$ Vector of observations
$X_{nxk} \rightarrow$ Matrix of observations
$\beta_{kx1} \rightarrow$ Vector of Coefficients
$\varepsilon_{nx1} \rightarrow$ Vector of Error terms

The Ordinary Least Square estimator is given by:

$$\beta = S^{-1}X'y, \ \ where\ S = X'X \tag{3}$$

Nwakuya/Nwabueze shrinkage estimator is given by;

$$\hat{b}_{N/Nse} = \left(1 - \frac{(m-2)\tau}{\beta' X' X \beta}\right)\beta \tag{4}$$

Where τ is defined as $\tau = \dfrac{\varepsilon'\varepsilon}{n-m}$, where $\varepsilon = y - X\beta$, n is the sample size and m is the number of imputations. This estimator can be seen in [2] as an extension of Ohtani's shrinkage estimator, which he proposed based on Stein's estimator [7].

The minimum mean square error (MMSE) estimator is given by;

$$\hat{b}_{MMSE} = \left(\frac{\beta' S\beta}{\beta' S\beta + \varepsilon'\varepsilon / \upsilon}\right)\beta \tag{5}$$

where $\upsilon = n - k$, k is defined as the number of parameters, while n is the sample size. This estimator is a biased estimator but was proved to be the best in terms of mean square error among the class of linear homogenous estimators [9].

The Adjusted Minimum Mean Square Error estimator is given by;

$$\hat{b}_{AMMSE} = \left(\frac{\beta' S\beta / k}{\beta' S\beta / k + \varepsilon'\varepsilon / n-k}\right)\beta \tag{6}$$

Where all parameters are as earlier defined in equations 3 and 4. This estimator was proposed by Ohtani as an improvement to minimum mean square error estimator, and it showed to have a smaller mean square error.

In this paper, we introduced number of imputation (m) into the existing adjusted minimum mean square error estimator. We replaced the number of parameters (k) by the number of multiple imputations (m).

Our proposed estimator is given by;

$$\hat{b}_{PAMMSE} = \frac{\beta' S\beta / m}{\beta' S\beta / m + \varepsilon'\varepsilon / n-m}$$

Where m is the number of imputations and the other parameter as defined in equations 3 and 4.

3. Analysis

Three different normally distributed regression data sets of sample sizes 20000, 8000 and 30each with 10% missing values were simulated in R. The missing points on the data sets were suitably imputed using 6 different imputation numbers; m= 5, 15, 20, 30, 40 and 50. Each data set was imputed using each of the imputation numbers and analyzed independently using each of the four methods. The mean square errors were calculated for each of the methods. A comparison test among the methods was also done at 0.05 level of significance.

Below are the tabulated results (table 1.1).

- OLS $\Rightarrow$ Ordinary Least Square
- MMSE $\Rightarrow$ Minimum Mean Square Error Estimator
- N/NSE $\Rightarrow$ Nwakuya/Nwabueze Shrinkage Estimator
- PAMMSE $\Rightarrow$ Proposed Adjusted Minimum Mean Square Error Estimator

4. Observations

From table 1.1, visually we can see that the value of the mean square error was highest with the OLS followed by the MMSE then the N/N Shrinkage Estimator and then the proposed adjusted minimum mean square error. Going further a paired comparison test was carried out to determine if the mean square errors from the four methods were significantly different from each other. Results shown in table 1.2 for the comparison test, shows that the mean square errors are significantly different from each other this shows that the visual differences seen among the estimators are statistically significant.

Table 1.1. Mean Square Error from Four Different Estimators with Different Sample Sizes and Different Imputation Numbers

Sample sizes	Number of imputations	OLS	MMSE	N/NSE	PAMMSE
n=20,000	m=5	48,194.37	48,194.27	48,194.18	48,193.98
	m=15	146,500.70	146,500.30	146,499.90	146,499.10
	m=20	196,411.80	196,411.20	196,410.60	196,409.50
	m=30	289,905.20	289,904.50	289,903.80	289,902.40
	m=40	385,379.10	385,378.00	385,377.10	385,375.40
	m=50	482,565.70	482,564.30	482,563.20	482,561.00
n=8,000	m=5	132,487.00	132,485.10	132,483.30	132,479.70
	m=15	391,895.70	391,890.50	391,885.40	391,875.20
	m=20	525,075.10	525,067.90	525,060.90	525,047.00
	m=30	763,493.30	763,484.70	763,476.40	763,459.90
	m=40	999,603.10	999,593.30	999,584.00	999,565.50
	m=50	1,237,682.00	1,237,660.00	1,237,660.00	1,237,639.00
n=30	m=5	50,808,245.00	50,766,908.00	50,726,301.0	50,646,284.0
	m=15	142,449,068.00	142,358,361.0	142,268,665.0	142,095,667.0
	m=20	190,790,183.00	190,692,661.0	190,597,371.0	190,406,647.0
	m=30	285,304,866.00	285,113,218.0	284,923,769.0	284,557,326.0
	m=40	384,361,812.00	384,143,649.0	383,929,013.0	383,507,811.0
	m=50	476,389,209.00	476,151,358.0	475,917,370.0	475,457,707.0

Table 1.2. Paired Comparison Test for the Four Estimators

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	OLS – MMSE	48738.167	83167.73883	19602.824	7379.82315	90096.51018	2.486	17	.024
Pair 2	OLS – N/NSE	96721.572	165071.70393	38907.774	14633.34453	178809.79881	2.486	17	.024
Pair 3	OLS – PAMMSE	190673.69	325362.8779	76688.766	28874.53567	352472.84100	2.486	17	.024
Pair 4	MMSE – N/NSE	47983.405	81904.20396	19305.006	7253.40259	88713.40741	2.486	17	.024
Pair 5	MMSE –PAMMSE	141935.522	242195.5243	57086.033	21494.52095	262376.52239	2.486	17	.024
Pair 6	N/NSE– PAMMSE	93952.12	160294.2773	37781.724	14239.64792	173664.58541	2.487	17	.024

5. Conclusions

The results have shown that indeed the proposed adjusted minimum mean square error has the least mean square error and at such dominates the other estimators. We can conclude that amongst the four estimators presented in this work the proposed minimum mean square error estimator seemed to do better than the other estimators considered in this research work.

REFERENCES

[1] En.m.wikipedia.org/wiki/Mean_square_error, Mean Square Error. Assessed 20th June 2016.

[2] 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 sciences, Vol 4, No 1, pp 65.

[3] Ohtani K (1996), 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.

[4] Royston P. (2004), Multiple imputation of missing values; The Stata journal, Vol 4 No 3, pp 227-241.

[5] Rubin D.B. (1976), Inference and missing data; Biometrika, Vol 63. No 3, pp 581-592.

[6] Rubin D.B. (1987), Multiple imputation for non-response in surveys, John Wiley and Sons, New York, pp 546-550.

[7] 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, No1, pp 197-206.

[8] Theil H. (1971), Principles of econometrics, New York: John Wiley.

[9] Wan A. T. K and Ohtani K. (1999), Minimum mean-square error estimation in linear regression with an inequality constraint; *Journal of Statistical Planning and Inference*, Vol 86, No 2000, pp 157-173.