Available online at www.elixirpublishers.com (Elixir International Journal)

International Business Management

Elixir Inter. Busi. Mgmt. 124 (2018) 52244-52247

ISSN: 2229-712X

EM Method of Estimation of Missing Data (Key Assumptions and Method for Applied Analysis)

Montasir A. A. Mohammed[1], Adil M. Y. Waniss[1] and Khalid R. K. Genawi[2]
[1]College of Business and Economics, Qassim University, Kingdom of Saudi Arabia.
[2]College of science, Sudan University of Science and Technology, Republic of Sudan.

ARTICLE INFO

Article history:
Received: 29 October 2018;
Received in revised form:
17 November 2018;
Accepted: 27 November 2018;

Keywords
Missing data,
Missing completely at random (MCAR),
Not missing completely at random (NMCAR),
Expectation Maximization (EM).

ABSTRACT

Missing data are pervasive, and pose problems for many statistical procedures. We all should be using methods that treat missing data properly, rather than deleting data or using single imputation. Importantly, it is not difficult to implement these missing data. A relatively few absent observations on some variables can dramatically shrink the sample size. As a result, the precision of confidence intervals is harmed, statistical power weakens and the parameter estimates may be biased. In this paper we used EM method that aims to estimate Missing Values, and summarizes how to apply out EM method to estimation missing data using SPSS.

Introduction

This paper is concerned with EM method of estimation missing data. missing data may be ignored or neglected, which may lead to Estimates that has less efficient, and may limit use of some statistical methods that require that there are no missing data, In this case, some bias may occur in results and weakness in power of statistical tests and measurements used.

Attention has to be taken about missing data, their processing and the mechanism of dealing with them. With the increase in development progresses in statistical programs that use computers to proceed, researcher should pay attention by using most appropriate analysis of his data, in order to arrive to conclusions that have more accurate parameters. To achieve this objective, an appropriate method of processing missing data must be chosen before starting the analysis.

Problem of the Study

It has been demonstrated repeatedly that missing data can have large effects on the results of a survey. Moreover, increasing the sample size without targeting nonresponse does nothing to reduce bias in missing data; Increasing the sample size may actually worsen the nonresponse bias, as the larger sample size may divert resources that could have been used to reduce or remedy the nonresponse, or it may result in less care in the data collection [1]. The main problem caused by nonresponse is potential bias.

Objectives of the study

The study aims at identifying the efficiency of EM method for estimating missing data, by comparing its estimates with parameters of real data.

Hypotheses of the study

The paper hypothesizes the followings: firstly, missing completely at random (MCAR). Secondly, main hypothesis, there is a statistically significant difference between means of parameters and estimated values. Lastly, the correlation between parameters and estimated values is significant.

Methodology of the study

In this study, a descriptive and analytical approach are used to determine the efficiency of EM method for estimating missing data according to the accurate scientific conditions to increase the accuracy of the estimates of the estimators. The paper is based on the applied side of generated data with mean 1000 and variance 10. It also uses SPSS Program as analytical tool to estimate and analyze data.

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 [2]. Accordingly, some studies have focused on handling the missing data problems caused by missing data, and the methods to avoid or minimize such in science researches [3, 4].

Missing data mechanisms

There are different assumptions about missing data mechanisms:

a) Missing completely at random (MCAR): Missing completely at random (MCAR): Suppose variable Y has some missing values. We will say that these values are MCAR if the probability of missing data on Y is unrelated to the value of Y itself or to the values of any other variable in the data set. However, it does allow for the possibility that “missingness” on Y is related to the “missingness” on some other variable X. [5] [6]

Tele:
E-mail address: montasirstat@gmail.com

b) Missing at random (MAR): a weaker assumption than MCAR: The probability of missing data on Y is unrelated to the value of Y after controlling for other variables in the analysis (say X). Formally: P(Y missing |Y, X) = P(Y missing |X) [6].

c) Missing not at random (MNAR): Finally, data are missing not at random (MNAR) when the probability of missing data on a variable Y is related to the values of Y itself, even after controlling for other variables [7]

The EM Algorithm

This algorithm is a parametric method to impute missing values based on the maximum likelihood estimation. This algorithm is very popular in statistical literatures and has been discussed intensively by many researchers, such as: [8], [9], [10], and [11].

This algorithm uses an iterative procedure to finding the maximum likelihood estimators of parameter vector through two step described in Dempsteret al. [9] and [10] as follows:

a). The Expectation step (E-step)

The E step is the stage of determining the conditional expected value of the full data of log likelihood function $l(\theta|Y)$ given observed data. Suppose for any incomplete data, the distribution of the complete data $\mathbf{Y}$ can be factored as

$$\mathbf{f(Y|\theta) = f(Y_{mis}\,, Y_{obs}\,|\theta)}$$
$$= \mathbf{f(Y_{obs}|\theta)\; f(Y_{mis}|Y_{obs}\,, \theta)} \qquad (1)$$

Where $\mathbf{f(Y_{obs}|\theta)}$ the distribution of the data is observed $\mathbf{Y_{obs}}$ and $\mathbf{f(Y_{mis}\,, Y_{obs}\,|\theta)}$ is the distribution of missing data given data observed. Based on the equation (1), we obtained log likelihood function

$$\mathbf{l(\theta|Y) = l(\theta|Y_{obs}) + \log f(Y_{mis}|Y_{obs}\,, \theta)} \qquad (2)$$

Where $l(\theta|Y)$ is log likelihood function of complete data, $l(\theta|Y_{obs})$ is log likelihood function of observed data, and $f(Y_{mis}|Y_{obs}\,, \theta)$ is the predictive distribution of missing data given $\boldsymbol{\theta}$

Objectively, to estimate $\boldsymbol{\theta}$ is done by maximizing the log likelihood function (2). Because $\mathbf{Y_{mis}}$ not known, the right side of equation (2) can not be calculated. As a solution, $\mathbf{l(\theta|Y)}$ is calculated based on the average value $\log f(Y_{mis}|Y_{obs}\,, \theta)$ using predictive distribution $f\big(Y_{mis}\big|Y_{obs}\,, \theta^{(t)}\big)$, where $\boldsymbol{\theta}^{(t)}$ is temporary estimation of unknown parameters. In this context, an initial estimation $\boldsymbol{\theta}^{(0)}$ be calculated using the complete case analysis. With this approach, the mean value of equation (4) can be expressed

$$\mathbf{Q\big(\theta|\theta^{(t)}\big)} =$$
$$\mathbf{l(\theta|Y_{obs}) + \int \log f(Y_{mis}|Y_{obs}\,, \theta)\, f\big(Y_{mis}\big|Y_{obs}\,, \theta^{(t)}\big)\, \partial Y_{mis}}$$
$$= \int \mathbf{[l(\theta|Y_{obs})}$$
$$+ \int \mathbf{\log f(Y_{mis}|Y_{obs}\,, \theta)]\;\; f\big(Y_{mis}\big|Y_{obs}\,, \theta^{(t)}\big)\, \partial Y_{mis}}$$
$$= \int \mathbf{l(\theta|Y)\; f\big(Y_{mis}\big|Y_{obs}\,, \theta^{(t)}\big)\, \partial Y_{mis}} \qquad (3)$$

The equation (3) basically a conditional expected value of log likelihood function for complete data $\mathbf{l(\theta|Y)}$ given observed data and initial estimate of unknown parameter.

b). the maximization step (M-step)

The M step is to obtained the iteratively estimation $\boldsymbol{\theta}^{(\mathbf{t+1})}$ with maximizes $Q\big(\theta\big|\theta^{(t)}\big)$ as follow

$$\mathbf{Q\big(\theta^{(t+1)}\big|\theta^{(t)}\big) \;\geq\; Q\big(\theta|\theta^{(t)}\big)} \qquad (4)$$

Both E and M steps are iterated until convergent.

Application

In this Section, we will introduce the applied side

Results and discussion

Firstly: Results related by the first case MCAR

i. With respect to Little's MCAR test

To test this hypothesis, we calculated values of Chi-Square and p-value, table (1) shows that

Table (1). little's MCAR test, Chi-Square and p-value.

missing data	Little's MCAR	
	Chi-Square	p-value
5%	0.122	0.727
10%	2.471	0.116
15%	0.547	0.460

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

The above table shows the p-values of little's MCAR test of the missing data of 5% , 10% and 15% are respectively (0.727), (0.116) and (0.460) which are greater than significant level (0.05). And, that means, the missing completely at random (MCAR).

ii. With respect to Std. Error Mean for real data and estimates EM method

We calculated means and std. deviation of deviation errors depend on parameters of real data and estimates of EM method. Table (2) shows that

Table (2). Comparison of Mean, Std. Deviation and Std. Error Mean results between parameters of real data and estimates EM.

missing data	real data			EM method		
	mean	Std. Deviation	Std. Error Mean	mean	Std. Deviation	Std. Error Mean
5%	1005.34	10.14	4.54	1002.32	1.34	0.60
10%	1006.79	6.45	2.05	1002.38	0.40	0.13
15%	1000.86	11.32	2.92	1002.15	1.77	0.46

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

From the above table, it has been shown that according to the mean parameters of real data, missing data of 10% have the highest mean (1006.79), followed by missing data of 5% depending on the value of the second largest mean (1005.34), lastly missing data of 15% have the less mean (1000.86). With std. deviation values, missing data of 10% have the lowest std. deviation (6.45), followed by missing data of 5% and missing data of 15% (10.14) and (11.32) respectively. And std. error mean values, missing data of 10% have the lowest std. error mean (2.05), followed by missing data of 5% and 10% are (2.92) and (4.54) respectively.

And mean values of estimates EM method, missing data of 10% have the highest mean (1002.38), followed by missing data of 5% depending on the value of the second largest mean (1002.32). Lastly, missing data of 15% have the less mean (1002.15). With std. deviation values, missing data of 10% have the lowest Std. deviation (0.40), followed by missing data 5% and missing data 15% (1.34) and (1.77) respectively. And std. error mean values, missing data of 10% have the lowest std. error mean (0.13), followed by missing data of 15% and 5% are (0.46) and (0.60) respectively.

From table (2), we notice that there is ostensibly difference between means of deviation errors for parameters of real data and EM data variables, and to know the statistical significance of differences, we used t-test. table (3) shows that

iii. With respect to t-test

To test this hypothesis, we calculated values of t-test and p-value, table (3) shows that

Table (3). t-test, p-value and Mean difference.

missing data	t-test		
	t	p-value	Mean difference
5%	0.661	0.527	3.02
10%	2.141	0.061	4.40
15%	0.436	0.669	1.28

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 5%, 10% and 15% are respectively (0.527), (0.061) and (0.669) which are greater than significant level (0.05). And, therefore, there is no a statistically significant difference between means. We concluded that to impute the missing values we can be used the EM method in case MCAR and missing data percentages 5%, 10% or 15%.

iv. With respect to Covariances and Correlations

We calculated Covariances and Correlations depending on parameters of real data and estimates EM method variables. table (4) shows that

Table (4). Covariances and Correlations, Pearson Correlation and p-value. Between parameters of real data and estimates EM method.

missing data	Covariances	Correlations	
		Pearson	p-value
5%	0.796	0.059	0.926
10%	-0.377	-0.146	0.687
15%	-2.255	-0.111	0.693

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

From the above table, it has been shown that according to the covariances values between parameters of real data and estimates EM method, missing data of 5% covariance is (0.796), missing data of 10% covariance is (-0.146), lastly missing data of 15% covariance is (-0.111). And correlations values between parameters of real data and estimates EM method, missing data of 5% pearson correlation is (0.059), which indicates that there is a moderate positive relationship, with p-value (0.926) that means, the correlation is not significant, missing data of 10% pearson correlation is (-0.146), which indicates that there is a very weak negative relationship, with p-value (0.687) that is means, the correlation is not significant, missing data of 15% pearson correlation is (-0.111), which indicates that there is a very weak negative relationship, with p-value (0.693) that is means, the correlation is not significant.

Secondly: Results related by the second case NMCAR

i. With respect to Little's MCAR test

To test this hypothesis, we calculated values of Chi-Square and p-value, table (5) shows that

Table(5). little's MCAR test, Chi-Square and p-value.

missing data	Little's MCAR	
	Chi-Square	p-value
5%	4.507	0.034
10%	5.970	0.015
15%	6.988	0.008

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

From the above table, it shows the p-values of little's MCAR test of the missing data of 5% , 10% and 15% are respectively (0.034), (0.015) and (0.008) which are less than significant level (0.05). And, that is means, not missing completely at random (NMCAR).

ii. With respect to Std. Error Mean for parameters of real data and estimates EM method

We calculated means and std. deviation of deviation errors depend on parameters of real data and estimates EM method. table (6) shows that

Table (6). Comparison of Mean, Std. Deviation and Std. Error Mean results between parameters of real data and estimates EM method.

missing data	real data			EM method		
	mean	Std. Deviation	Std. Error Mean	mean	Std. Deviation	Std. Error Mean
5%	1003.98	7.00	3.13	1001.17	3.26	1.46
10%	1008.84	9.93	3.14	1001.27	10.82	3.42
15%	995.44	8.22	2.12	1003.37	0.18	0.05

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

From the above table, it has been shown that according to the mean parameters of real data, missing data of 10% have the highest mean (1008.84), followed by missing data of 5% depending on the value of the second largest mean (1003.98), lastly missing data of 15% have the less mean (995.44). With std. deviation values, missing data of 5% have the lowest std. deviation (7.00), followed by missing data of 15% and 10% are (8.22) and (9.93) respectively. And std. error mean values, missing data of 15% have the lowest std. error mean (2012), followed by missing data of 5% and 10% are (3.13) and (3.14) respectively.

And mean values of estimates EM method, missing data of 15% have the highest mean (1003.37), followed by missing data of 10% depending on the value of the second largest mean (1001.27), lastly missing data of 5% have the less mean (1001.17). With std. deviation values, missing data of 10% have the lowest std. deviation (0.18), followed by missing data of 5% and missing data of 10% are (3.26) and (10.82) respectively. And std. error mean values, missing data of 15% have the lowest std. error mean (0.05), followed by missing data of 5% and missing data of 10% are (1.46) and (3.43) respectively.

We note from table (6) there is ostensibly difference between means of deviation errors for parameters of real data and EM data variables, and to know the statistical significance of differences, we used t-test. Table (7) shows that

iii. With respect to t test

To test this hypothesis, we calculated values of t-test and p-value, table (7) shows that

Table (7). t-test, p-value and Mean difference.

missing data	t-test		
	t	p-value	Mean difference
5%	1.102	0.315	3.80
10%	1.630	0.120	7.57
15%	3.737	0.002	7.93

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

From the above table, it shows the p-values of t-test for the missing data 5%, 10% are respectively (0.315), (0.120) are greater than significant level (0.05). And, therefore, there is no a statistically significant difference between means. p-value of t-test for the missing data 15% (0.002) is less than significant level (0.05). And, therefore, there is a statistically significant difference between means. Hence, based on those results, we concluded that to impute the missing values we can used the EM method in case NMCAR and missing data percentages 5% or 10%.

iv. With respect to covariances and correlations

We calculated covariances and correlations depend on real data and estimates estimates EM method. table (8) shows that

Table (8). Covariances and Correlations, Pearson Correlation and p-value. Between real data and estimates EM method.

missing data	Covariances	Correlations	
		Pearson	p-value
5%	-15.60	-0.683	0.204
10%	18.19	0.169	0.640
15%	0.490	0.326	0.236

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

From the above table, it has been shown that according to the covariances values between parameters of real data and estimates EM method, missing data of 5% covariance is (-15.60), missing data of 10% covariance is (18.19), lastly missing data of 15% covariance is (0.490). And correlations values between parameters of real data and estimates EM method, missing data of 5% pearson correlation is (-0.683), which indicates that there is a moderate negative relationship, with p-value (0.204) that is means, the correlation is not significant, missing data of 10% pearson correlation is (0.169), which indicates that there is a very weak positive relationship, with p-value (0.640) that is means, the correlation is not significant, missing data of 15% pearson correlation is (0.326), which indicates that there is a weak positive relationship, with p-value (0.236) that is means, the correlation is not significant.

References

[1]Sharon L. Lohr 2010 Sampling: Design and Analysis, Second Edition, Arizona State University, 330-323

[2]Graham JW. 2009; Missing data analysis: making it work in the real world. Annu Rev Psychol. 60:549-576.

[3]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

[4]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.

[5]Briggs, A., Clark, T., Wolstenholme, J., Clarke, P., 2003. Missing presumed at random: cost-analysis of incomplete data, Health Economics 12, 377-392

[6]Allison, P., 2001. Missing data-Quantitative applications in the social sciences. Thousand Oaks, CA: Sage. Vol. 136.

[7]Craig K. Enders. 2010; Applied Missing Data Analysis. Guilford Press - Psychology - 377 pages

[8] Little R J A and Rubin D B 1987 Statistical Analysis with Missing Data (New York: John Wiley & Son Inc.)

[9]Dempster A P, Laird N M, and Rubin D B 1977 Journal of the Royal Statistical Society Series B 39 (1) 1-38

[10]Little R J A and Rubin D B 2002 Statistical Analysis with Missing Data Second Edition (Hoboken, New Jersey: John Wiley & Son Inc.)

[11]Watanabe M and Yamaguchi K 2004 The EM Algorithm and Related Statistical Models (New York: Marcel Dekker, Inc.)