JOURNAL OF SCIENCE

SAKARYA UNIVERSITY

Sakarya University Journal of Science

ISSN 1301-4048 | e-ISSN 2147-835X | Period Bimonthly | Founded: 1997 | Publisher Sakarya University | http://www.saujs.sakarya.edu.tr/

Title: Investigation of the Multiple Imputation Method in Different Missing Ratios and Sample Sizes

Authors: Nesrin Alkan, B. Baris Alkan
Recieved: 2019-01-03 00:00:00

Accepted: 2019-02-05 00:00:00

Article Type: Research Article
Volume: 23
Issue: 4
Month: August
Year: 2019
Pages: 605-609

How to cite
Nesrin Alkan, B. Baris Alkan; (2019), Investigation of the Multiple Imputation Method in Different Missing Ratios and Sample Sizes. Sakarya University Journal of Science, 23(4), 605-609, DOI: 10.16984/saufenbilder.507450
Access link
http://www.saujs.sakarya.edu.tr/issue/43328/507450

Sakarya University Journal of Science 23(4), 605-609, 2019

Investigation of the Multiple Imputation Method in Different Missing Ratios and Sample Sizes

Nesrin Alkan*[1], B. Baris Alkan[2]

Abstract

In many studies, missing data are the real trouble to researchers. Because the statistical methods are designed for complete data sets. Multiple imputation method is developed to solve the missing data problem. The method is also used effectively in some useful properties of the Bayes method. If there are missing values in the data set, Bayesian method can be used to prevent the loss of information. In this study, the performance of the multiple imputation method is evaluated by generating survival data with different missing rates and different sample sizes. Also, informative priors and multiple imputation method are used together to prevent the missing information in the variable with missing value.

Keywords: Missing Value, Multiple Imputation, Bayesian Cox Regression, Cox regression

1. INTRODUCTION

In many studies, researchers are interested in the comparison of the different groups. Observations in groups may have many features. We can give examples such as demographic variables, physiological variables, behavioral variables. These variables are called covariate or independent variable. Cox regression analysis is a method which is used to determine the cause-and-effect relationship between dependent variable and covariate. Missing data are a real disaster to researchers in many of disciplines. Because conventional statistical methods and softwares presume that all variables are measured for all observation. For this reason, the missing data problem must be solved. To deal with missing data problem, there are two ways that are removing the observation with missing value which is called complete case analysis or filling in the missing values. In Complete Case Analysis, observation with any missing value on the variable is deleted. Filling the missing values that is called imputation method yields a complete data set. So it is an attractive strategy. One of the most important imputation methods is multiple imputation method. In multiple imputation method, a few copies of the data set are produced and through the use of different estimates of missing values are filled in each copy. On the other hand, in multiple imputation each of missing values are filled in m (m=3 or m=5) times for generating *m* complete

[*] Corresponding Arthur: nesrinalkan@sinop.edu.tr

[1] Sinop University, Department of Statistics, Sinop, Turkey. ORCID:0000-0003-1452-4780

[2] Akdeniz University, Department of Educational Sciences, Antalya, Turkey. bbalkan@akdeniz.edu.tr ORCID:0000-0002-5851-7833

data sets. Each of the imputed data sets is analyzed by any statistical method and then the results are combined to make inference [1].

In this study, simulated data sets are used to determine the effect of different sample size and different missing value rates on multiple imputation method. For this reason, sample size are determined N=50, 100, 200 and missing rate is determined 5%, 10%, 20%, 40%. Multiple imputation method is applied for each of the data sets. Results of the multiple imputation compare with the original results. As real data set, lung cancer patient data set with missing value is used. Firstly, missing data problem is solved with multiple imputation metod and then informative priors are used in Bayesian Cox regression to determine important prognostic factors in order to prevent the loss of information that missing values cause. Informative priors are obtained from a similar previous study. For application, both simulated data and a real data set are used.

2. BAYESIAN COX REGRESSION ANALYSIS

Cox regression analysis is a pupular method in survival analysis. The Cox model is written as following equation

$$h(t;x_i)=h_0(t)\exp(\beta' x_i) \quad (1)$$

where β is the unknown parameter vector, x is the covariate vector, and $h_0(t)$ is called the baseline hazard h(t,x) denotes the hazard function [2]. In Bayesian methods, researchers can make probability statements about parameters. Thomas Bayes in 1763 formulated Bayes theorem and the basis of the Bayesian approach is obtained from this theorem. There are three basic elements in Bayesian analysis. Firstly a prior distribution is determined. A prior distribution is information about the parameter. Secondly, the likelihood function is used to summarize the data. Finally the posterior distribution is calculated. In the posterior distribution, information from the the prior distribution and likelihood function are combined. So the information about the parameter θ is updated [3]. The posterior distribution for Bayesian Cox Regression is formulated as following equation

$$P(\beta/x) \propto L(\beta)P(\beta) \quad (2)$$

where L(β) is the partial likelihood function. If the posterior distribution is known, the posterior point estimate can be found. Prior distribution has a very important task in Bayesian Cox regression. Generally a normal prior is used to choose the informative prior and a uniform prior is used to select a non-informative prior for β. Informative priors provide access to previous studies measuring the same response and covariates as the current study [4].

3. MULTIPLE IMPUTATION METHOD

Researchers encounter with missing in many areas such as social, behavioral and medical sciences. Missing data is a big problem as classic statistical methods and softwares which need to provide some important assumption. Multiple imputation method (MI) uses Bayesian techniques to deal with the missing data problem. A multiple imputation method consists of imputation, the analysis and pooling phase. In the imputation phase, imputed data sets are obtained (e.g., m=5). Each of data sets contains different estimates of the missing values [5], [6].

The imputation phase of multiple imputation method contains two-step that consists of imputed step (I-step) and a posterior step (P-step). In the imputed step, covariance matrix and mean vector are estimated and regression equations are created to estimate the missing values of covariates from the observed data. In the P step, alternate estimates of covariance matrix and mean vector are produced. After P-step the new estimates are used in I-step [1]. In the analysis phase, standard statistical methods are used for each imputed data sets. Pooling phase is the last step. In this step, the results obtained from these methods are combined for the inference [7], [8].

4. APPLICATIONS

Simulation study

In order to determine the effect of different sample sizes and different missing value rates on multiple imputation method, simulations are made. In the simulation study, as first sample size N = 50, 100, 200 are selected. Each of data with different sample size is analyzed by Cox regression analysis. Regression parameters and standard errors are estimated. These estimates are called as true value (0% missing rate). After this, some of the values of the variables are deleted in different rates to obtain incomplete data sets which have 5%, 10%, 20% and 40% missing value. Each of data sets with different missing rate is analyzed by Cox regression, and then parameters are estimated and compared with true values. Each of the absolute error (0% missing rate) is found by calculating the difference between true value and estimated value for different missing rate. The results are given in Table 1.

Table 1: The results of different sample sizes and different missing value rates on Multiple Imputation Method

N=50				**Absolute Error**		
Missing rates	X1	X2	X3	X1	X2	X3
0%	-0.01	-0.81	0.10	-	-	-
5%	-0.03	-0.87	0.10	0.020	0.06	0.00
10%	0.025	-0.88	0.12	0.035	0.07	0.02
20%	0.029	-0.96	0.13	0.039	0.15	0.03
40%	0.200	-1.08	0.16	0.210	0.27	0.06
N=100				**Absolute Error**		
0%	-0.14	-0.31	0.104	-	-	-
5%	-0.15	-0.32	0.107	0.01	0.01	0.003
10%	-0.13	-0.35	0.094	0.01	0.04	0.010
20%	-0.10	-0.37	0.078	0.04	0.06	0.026
40%	-0.06	-0.38	0.077	0.08	0.07	0.027
N=200				**Absolute Error**		
0%	-0,129	-0.233	0.077	-	-	-
5%	-0.13	-0.23	0.08	0.001	0.003	0.003
10%	-0.12	-0.22	0.07	0.009	0.013	0.007
20%	-0.11	-0.22	0.064	0.019	0.013	0.013
40%	-0.10	-0.19	0.069	0.029	0.043	0.008

According to the Table 1, in all sample sizes and different missing rates, multiple imputation method yields estimates close to real value so the absolute error is close the zero. The estimates are close to the actual parameter regardless of the loss rate for N = 200. However, in case of N = 50, as the missing rate increases, the multiple imputation method is less successful. According to this result, the missing rate is important in small samples but not in large samples. As the sample size grows, the results are close to real values no matter what the missing rate is.

Real data set

At this stage of the study the survival data of patients with lung cancer is obtained from the Faculty of Medicine in Ondokuz Mayis University. In the data, some observations of variables are deleted to generate missing data with 20% missing values. In this study, survival data with missing value is completed with multiple imputation method and parameter estimates are made by applying Cox regression and Bayesian Cox regression analysis respectively. Cox regression analysis is applied for lung cancer data set (original complete data) to obtain true value. Then Cox regression and Bayesian Cox regression are applied for lung cancer data set which is imputed with multiple imputation method and obtained estimations of parameters. The estimates obtained separately from the missing data, Cox regression applied after multiple imputation method and Bayesian Cox regression applied after multiple imputation method are summarized in Table 2. In Bayesian Cox Regression, informative priors are used. For informative priors, previous studies related to lung cancer were examined. Hazard ratios as informative priors for hemoglobin, protein and LDH levels are respectively 2.2, 1.3, 0.8 in [9], for albumin, hazard ratio is 0.95 in [4] and for tumor size, hazard ratio is 1.08 in [10].

In this study, we also study to determine the method which obtains the closest value to the regression coefficient of the original data. For this reason, difference between true value (from original complete data) and estimation of Cox regression after multiple imputation and Bayesian Cox regression after multiple imputation are examined and absolute errors are computed and results are given Table 3.

Table 2: Summary results of the analysis

Parameter	Original data	Multiple Imputation +Cox	Multiple Imputation +Bayesian Cox
Weight loss	-0,1290	-0,1355	-0,0737
Hemoglbn	0.1890	0.1761	0.1785
Platelet	8.31E-07	6.70E-07	8.23E-07
Protein	-0.3605	-0.3177	-0.3539
Albumin	-0.1195	-0.1384	-0.1128
LDH	0.0002	0.0001	0.0002
ECOG	1.0849	1.1186	0.9987
Hist. type	-0.2331	-0.2434	-0.2326
Tumorsize	0.0769	0.0851	0.0695

Table 3: Absolute Error of estimation

	Absolute Error	
Parameter	Multiple Imputation +Cox	Multiple Imputation +Bayesian Cox
Weight loss	0.0065	0.0553
Hemoglbn	0.0129	0.0105
Platelet	1.61E-07	8E-09
Protein	0.0428	0.0066
Albumin	0.0189	0.0067
LDH	0.0001	0.00001
ECOG	0.0337	0.0862
Hist. type	0.0103	0.0005
Tumorsize	0.0082	0.0074

According to the Table 3, it has been seen that multiple imputation method solves the missing data problem because both Cox regression and Bayesian Cox regression obtain very close estimation to original data. Additionally Bayesian Cox regression has smaller absolute error than Cox regression for variables which have informative prior. Bayesian Cox regression with informative priors give values the closest to the regression coefficients which are obtained from the original data for hemoglobin, platelet, protein, albumin, LDH, histological type, tumor size.

5. CONCLUSION

Missing data is that some of the values of variables are missing. Many researchers are throwing out the observation with any missing value. However, it leads to loss of information, decreases the sample size and statistical power. For this reason, in missing data problem, the use of imputation methods is suggested. Multiple imputation method is one of the imputation method that performs better than other methods [11]. In missing data problem, for the parameter estimation, very little information can be obtained from the data and using informative priors can help to overcome this problem [4]. Both the method using informative priors for variables which have missing values and multiple imputation method have been used together to prevent the decrease in data information and to find the method that gives the most accurate estimation.

In this study, survival data sets with different missing rate and different sample size are produce to determine the performance of multiple imputation. We have found that the missing rate is important in small samples and in large samples missing rate is not important. Another finding from the study is that if the sample size grows, parameter estimates are close to real values no matter what the missing rate is. As a result, multiple imputation method has been successful in solving the missing data problem. Better results are also obtained when using Bayesian methods with informative priors after multiple imputation method.

6. REFERENCES

[1] C. K. Enders, “Applied Missing Data Analysis,” New York: Guilford Press, pp. 165–286, 2010.

[2] D. R., Cox, “Regression models and life tables,” Journal of the Royal Statistical Society 34, pp. 187–220, 1972.

[3] N. Alkan, “Assessing Convergence Diagnostic Tests for Bayesian Cox Regression,” Communication in Statistics-Simulation and Computation, Vol.46, No.4, pp. 3201-3212, 2017.

[4] J. G., Ibrahim, M. H., Chen, D. Sinha, "Bayesian Survival Analysis. New York," Springer-Verlag, 2001.

[5] P. D. Allison, "Multiple imputation for missing data: a cautionary tale," Sociological Methods and Research 28 pp. 301–309, 2000.

[6] D. B. Rubin, "Inference and missing data," Biometrika 63 pp. 581–592, 1976.

[7] D. B. Rubin, "Multiple Imputation for Nonresponse in Surveys," 1st ed. New York, John Wiley&Sons, p. 303, 1987.

[8] J.L. Schafer, M.K. Olsen, "Multiple imputation for multivariate missing data problems: a data analyst's perspective," Multivariate Behavioural Research, vol. 33, no. 1, pp. 545-71, 1998.

[9] P.T., Lam, M.W., Leung, C.Y. Tse, "Identifying prognostic factors for survival in advanced cancer patients: a prospectivestudy," Hong Kong Med J, vol.13, pp. 453-459, 2007.

[10] C. M., Abreu, J. M., Chatkin, C. C., Fritscher, M. B., Wagner, J. A. L. F. Pinto, "Long-term survival in lung cancer after surgical treatment: is gender a prognostic factor?," 2003. (http://www.scielo.br/pdf/jbpneu/v30n1/en_v30n1a03.pdf. 26.06.2018)

[11] N., Alkan, Y., Terzi, M. A., Cengiz, B. B., Alkan, "Comparison of Missing Data Analysis Methods in Cox Proportional Hazard Models," Turkiye Klinikleri Journal of Biostatistic, vol. 5, no. 2, pp. 49-54, 2013.