Comparison of Missing Data Analysis Methods in Cox Proportional Hazard Models

Cox Oransal Hazard Modelinde Kayıp Veri Analizi Yöntemlerinin Karşılaştırılması

Nesrin ALKAN,[a]
Yüksel TERZİ,[b]
M. Ali CENGİZ,[b]
B. Barış ALKAN[a]

[a]Department of Statistics,
Sinop University
Faculty of Sciences and Arts, Sinop
[b]Department of Statistics,
Ondokuz Mayıs University
Faculty of Sciences and Arts, Samsun

Geliş Tarihi/*Received:* 09.01.2013
Kabul Tarihi/*Accepted:* 29.04.2013

This study is the re-reviewed form of the presentation titled "Comparison of Methods for Handling Missing Data in Cox Regression Models" which was presented in the "13th International Conference on Econometrics, Operations Research and Statistics-ICEOS 2012, Famagusta, North Cyprus, 24-26 May 2012."

Yazışma Adresi/*Correspondence:*
Nesrin ALKAN
Sinop University
Faculty of Science and Arts,
Department of Statistics, Sinop,
TÜRKİYE/TURKEY
nesrinalkan@sinop.edu.tr

ABSTRACT Objective: Data with missing value are common in clinical studies. This study investigated to assess the effects of different missing data analysis techniques on the performance of Cox proportional hazard model. **Material and Methods:** In order to see how sample size and missing rate effect the missing data analysis techniques, we derived the survival data with 25, 50 and 100 sample sizes. Some elements of the survival data with different sample size were deleted in different rates under MAR (Missing at Random) assumption to generate incomplete data sets which had 5%, 10%, 20% and 40% missing value for each data. Data sets with missing values were completed by five missing data analysis techniques (complete case Analysis-CCA, mean imputation, regression imputation-REG, expectation maximization-EM algorithm, multiple imputation-MI). The new completed data sets were analyzed by Cox proportional hazard model and their results were compared with results of original data. **Results:** The difference between the techniques grew for increasing missing rate and while the sample size increased the methods were similar to each other. CCA was the most affected from sample size. The estimates from the methods REG, EM and MI were very similar to each other and real value. **Conclusion:** Multiple imputation method as impute more than one value for each missing value should be preferred instead of single imputation methods as impute only one value for each missing value.

Key Words: Missing data; missing data analysis; Cox proportional hazard models; multiple imputation

ÖZET Amaç: Klinik çalışmalarda kayıp değerli verilerilerle çoksık karşılaşılır. Bu çalışmada kayıp değer problemini gideren yöntemlerin, Cox oransal hazard modelinin performansı üzerindeki etkisi incelendi. **Gereç ve Yöntemler:** Kayıp veri problemini gideren yöntemlerin örnek genişliğinden ve kayıp oranı miktarından nasıl etkilendiğini görmek amacıyla 25, 50 ve 100 birimlik sağkalım verisi türetilerek her bir veri setinde kayıp oranları %5, %10, %20 ve %40 olacak sekilde tesadüfi olarak kayıp-MAR varsayımına uygun kayıp değerli veri setleri oluşturuldu. Kayıp veri problemini gideren eksiksiz veri analizi-CCA, beklenti maksimizasyonu-EM, regresyon değer atama-REG, ortalama değer atama ve çoklu değer atama-MI yöntemleri oluşturulan kayıp değerli veri setlerine ayrı ayrı uygulandı ve performansları Cox oransal hazard modeli uygulanarak karşılaştırıldı. **Bulgular:** Kayıp oranı arttıkça Yöntemler arasındaki farkların büyüdüğü, buna karşın örnek genişliği arttıkça yöntemlerin birbirine benzediği görüldü. Örnek genişliğinden en fazla etkilenen yöntem CCA olarak belirlendi. Yöntemler içinde yaptıkları tahminler açısından EM, REG ve MI yöntemleri birbirine yakın sonuçlar verdi. **Sonuç:** Kayıp veri durumunda her bir kayıp değere birden fazla değer atayan çoklu değer atama yöntemi sadece bir değer atayan tekil değer atama yöntemlerine gore tercih edilebilir.

Anahtar Kelimeler: Kayıp veri; kayıp veri analizi; Cox oransal hazard modeli; çoklu değer atama

Turkiye Klinikleri J Biostat 2013;5(2):49-54

In clinical and epidemiological studies, researchers are often interested in the comparison of the different treatment groups. Individuals in groups may have additional features. For example, individuals may have many features, such as demographic variables (age, gender, etc.), physiological variables (blood glucose levels, blood pressure, etc.), behavioural variables (diet, smoking status, etc.). Such variables are called the independent variable or covariate and these variables are used to explain the dependent variable. Cox proportional hazard model of the most widely used method for modelling this types of data. The survival data have censored observations. For example in a clinical study the survival times for these patients are unknown so this observation are called censored.[1]

In the most of the survival data, individuals have variables with missing value. This kind of data is missing data. Complete case analysis is one of the solutions of problems of missing value in Cox proportional hazard models. This method is deleted cases with any missing value on the variable. [2] However complete case analysis may obtain ineffective results especially it deletes a large fraction of the sample. [3]

For this reason, deal with missing data problem and use of the method which overcomes missing data problems and impute closest estimations to the actual value instead of missing value is extremely important. Methods which enable the statistical analysis by solving missing data problem are called missing data analysis.

This study investigated to assess the effects of different missing data analysis techniques on the performance of Cox proportional hazard model.

MATERIAL AND METHODS

The Cox proportional hazard model is the most widely used method of survival analysis. In survival analysis, the Cox proportional hazard model is used to determine relation between dependent variable and covariates. The Cox proportional hazard model may be written as[1]

$$\lambda(t;\mathbf{z})=\exp(\mathbf{z}\boldsymbol{\beta})\,\lambda_0(t) \qquad (1)$$

where $\mathbf{z}$ is the covariate vector, $\boldsymbol{\beta}$ is the unknown parameter vector and $\lambda_0(t)$ is called the baseline hazard and it is function which obtain non parametric estimates as time-dependent and independent from $\mathbf{z}$'s.[4] $\lambda(t,\mathbf{z})$ represents the resultant hazard, given the values of the covariates for the situation with regard to survival time (t).

MISSING DATA ANALYSIS

Missing data often arise in various areas, especially in clinical trials, epidemiological studies. Here, the meaning of missing data that some of the values of variables are missing.

The missing data mechanisms are categorized for missing data problems by Rubin (1976) and Little&Rubin (2002). This mechanism describes the relationships between the missing value and the data. In general there are three types of missing data mechanism in the literature. These are called as Missing at Random (MAR), Missing Completely at Random (MCAR) and Missing Not at Random (MNAR).[5,6] If the probability of missing value on a variable depends on measurements of the other variables in analysis of the model but it is not connected to values of the variable, missing data mechanism is called MAR. If probability of missing value on a variable is not depended on measurements of the other variables and values of itself, data is called MCAR.[7]When the data is MNAR, probability of missing value on a variable depends on the values of itself but it does not depend on the others.[7]

Data sets with missing value are an important problem for researcher. Because statistical methods and software suppose that all variables in a model were measured for all cases. For this reason, the problem of data with missing values must be resolved.

There are two ways deal with missing data, that are removing the cases with missing value (Case Deletion) or filling in the missing values (Imputation Methods).[8]

Complete case analysis (CCA) is one of the methods most commonly used to resolve the missing data problems. Estimation of missing values by using known values of variables referred to as imputation. The most commonly used imputation methods are mean imputation, regression imputation, expectation maximization (EM) and multiple imputation (MI). One of the classical statistical analysis techniques may be applied to the new data which is completed with imputation methods.

Complete case analysis is known as listwise deletion. This method is deleted cases with any missing value on the variable. Despite the development of many methods that can complete the missing value, researchers often resort to this method in terms of easy to apply. Under MCAR assumption CCA method may not introduce bias. However this method will yield approximately unbiased estimates of Cox regression coefficients.[3]

In mean imputation, the missing values are filled with the arithmetic mean of the available observations for any variable in data matrix.[9]

Regression imputation establishes the regression equations to predict the missing value in variables from complete variables. This method used for many years is similar to mean imputation. The first step of the regression imputation method is to obtain regression equations which predict variables with missing values from complete variables and second step finds estimation of variables with missing values. These estimated values are used to replace missing values and the data set is completed.[7]

EM algorithm is a general method which helps to find maximum likelihood estimator in data with missing value. EM algorithm is a two-step procedure and provides good estimates under the assumption of multivariate normal distribution. The first step is called expectation (E) step which estimates the expectation of the logarithmic likelihood using observed value for the parameters. The second step is called maximization (M) step, a computed parameter estimates by maximizing the expected log-likelihood. These steps are repeated until convergence is achieved. EM algorithm can be described as iterated regression imputation. As first give the initial values for mean vector and covariance matrix. The E step uses the elements in the mean vector and the covariance matrix to obtain regression equations for predicting the missing value from the observed variables. The M step uses the real and imputed data to generate updated estimates of mean vector and covariance matrix. The updated parameter estimates forward to the next E step, these steps repeats until elements of mean vector and covariance matrix no more change.[10, 8, 5]

Mean, Regression, EM imputation methods generate a single replacement value for each missing data point. So they have called single imputation method. Most single imputation methods produce unbiased estimates under MAR assumption.[7]

Multiple imputation (MI) method develops the Bayesian approaches to solve the problem of missing value in the data. In multiple imputation each of missing values are filled in m times to generate m complete data sets. The imputation phase of multiple imputation method is two-step that consists of I-step and P-step. I- step use regression equations to predict the missing value of variables from observed data and add random residuals to the predicted value. In P-step, a new mean vector and a new covariance matrix are drawn randomly from their posterior distributions. After P-step, I-step uses the new estimate to obtain regression coefficient and different set of imputations. The imputed data sets are analyzed by standard statistical analysis and then combining the results from these analyses for the inference.[7] Rubin (1987) summarized the formulas for combined parameter estimates and standard errors. For example combined parameter estimates are arithmetic mean of the m complete data estimates. One of the basic decisions in multiple imputation method is to determine the number of imputed data sets. Rubin (1987) and Schafer (1998) recommend that three-five data sets are usually sufficient to get parameter estimation. Multiple imputation produces unbiased estimates if the data with missing value has MAR mechanism.[7,11,12,8,13]

RESULTS

This study was performed to assess the effects of different missing data analysis methods on the performance of Cox proportional hazard model. The most commonly used missing data analysis methods examined for different missing rates and different sample size with Cox proportional hazard model. For this purpose 25, 50, 100 survival data were used and some elements of these data sets were deleted in different rates under MAR (Missing at Random) assumption to generate incomplete data sets which had 5%, 10%, 20% and 40% missing value. Data sets with missing values were completed by five missing data analysis methods (complete case Analysis-CCA, mean imputation-MEAN, regression imputation-REG, expectation maximization algorithm-EM, multiple imputation-MI). The new complete data sets were analyzed by Cox proportional hazard model and their results were compared original complete data's results. The outcomes of interest were the regression coefficients and standard errors of covariates in the regression model. When comparing these missing data analysis, they examined in terms of closeness the real results which are obtained from data without missing data (original data).

Missing data analysis was applied for data with different missing rate when the sample size was 25. The completed data sets were analyzed with Cox proportional hazard models. The results of Cox proportional hazard models for completed data sets were compared with results of original data (Figure 1).

In Figure 1, all of the missing data analysis method yielded results close to the true regression coefficient when the missing rate was %5 and %10 for sample size of 25 units. With the increase in the rate of missing data, regression coefficients which obtained after CCA quite diverged from the true value of regression coefficients (Figure 1).

In general, MEAN method didn't give much run away from true parameters value and the best estimators. MEAN method showed a better performance compared with CCA. REG and MI generally gave the closest value to the true regression coefficient in 20% and 40% missing rate (Figure 1).

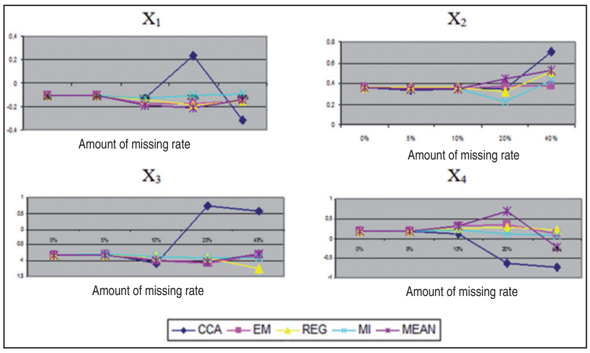

FIGURE 1: Regression coefficient estimates for different missing data analysis methods for different missing rate when sample size is 25 units.

Missing data analysis was applied to data with different missing rate when the sample size was 50. The completed data sets were analyzed with Cox proportional hazard models. When sample size is 50 units, the results of completed data sets were compared with results of original data (Figure 2).

In Figure 2 shown that in sample of 50 units CCA method gave better results than sample of 25 units. In this sample size all of the missing data analysis method yielded results close to the true regression coefficient when the missing rate was %5 and %10. EM and MI methods gave the closest value to the true regression coefficient in 20% and 40% missing rate. Estimates obtained from REG method were better than CCA and MEAN methods (Figure 2).

Missing data analysis was applied to data with different missing rate when the sample size was 100. The completed data sets were analyzed with Cox proportional hazard models. The results were compared with findings of original data (Figure 3).

According to Figure 3, CCA method yielded estimates close to the true value when missing rate is 20% and lower. After CCA method estimates diverged from the true value in 40% missing rate. In generally all of the missing data analysis method yielded results close to the true regression coefficient in lower 20% missing rate. Whereas EM, MI and REG methods gave the best estimates when missing rate was 40% (Figure 3).

We aimed to examine similarities of missing data analysis methods for increasing sample size. For this purpose the graphs were examined separately for all variables and similar results were obtained each of variables. So the graphs were obtained for one of the variables (X_1) (Figure 4).

The graphs which shown in Figure 4 were examined and for all sample size difference between missing data analysis methods were small in 5% missing rate. Especially when the sample size was 100 units, all missing data analysis methods gave almost the same estimates in 5% missing rate.

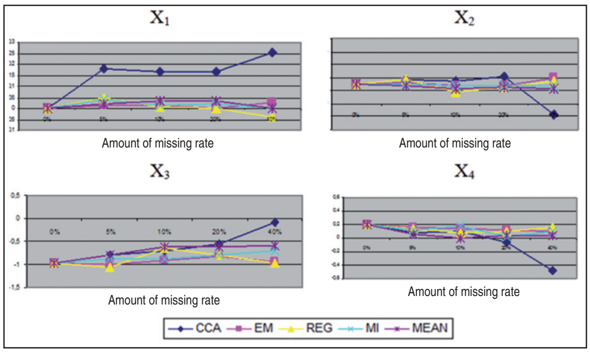

FIGURE 2: Regression coefficient estimates for different missing data analysis methods for different missing rate when sample size is 50 units.

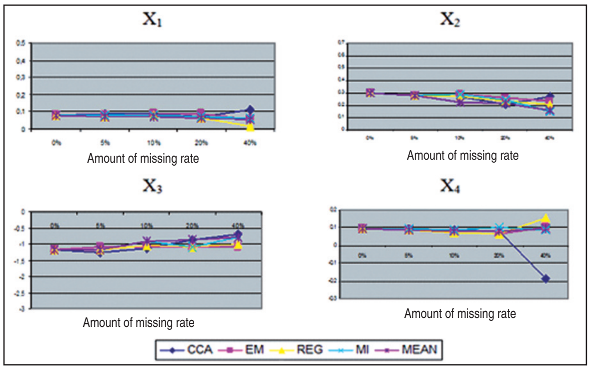

FIGURE 3: Regression coefficient estimates for different missing data analysis methods for different missing rate when sample size is 100 units.

Methods varied slightly in terms of their estimates in 10% missing rate for sample of 25 and 50 units. In this missing rate, methods obtained very similar results for sample of 100 units. In 20% missing rate, methods differed from each other more than 10% missing rate for sample of 25 and 50 units. Whereas all methods gave almost the same estimates for sample of 100 units. In 40% missing rate, methods differed from each other more than other missing rate (Figure 4).

Graphs of standard error for variable X_1 and X_2 in different sample size and different missing rate were shown (Figure 5).

In Figure 5 shown that applying missing data analysis methods expect CCA didn't affect the standard error for original data. But estimates after performing CCA were considerably increased when the missing rate increased (Figure 5). This is the result of reduction in sample size.

DISCUSSION AND CONCLUSION

In the literature, different imputation methods were compared. As a result they have seen multiple imputation method is often used in Cox proportional hazard model with missing covariate.[14] When the missing rate was 5%, there was very little difference between the missing data analysis. Also multiple imputation method can be preferred when the missing rate was over 10%.[15] But in our study we have discussed the five missing data analysis method there has not been before.

As a result of the experimental study, if the sample size is less than 50 (N<50), using this

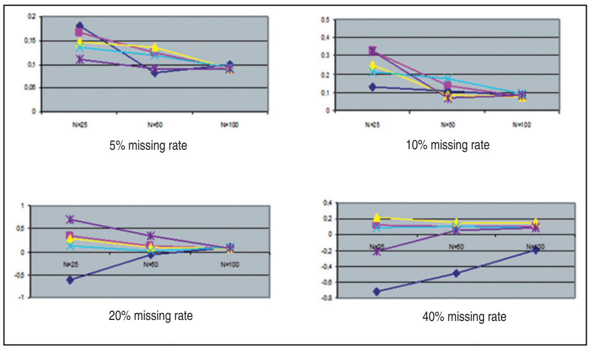

FIGURE 4: Missing data analysis methods for different missing rate for variable X_1.

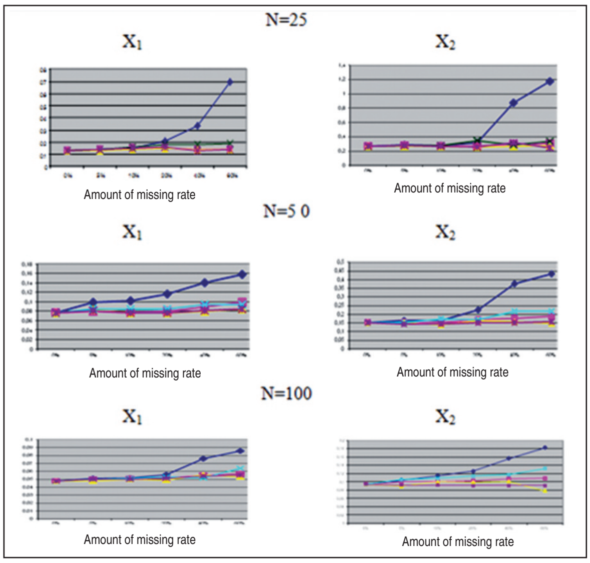

FIGURE 5: Standard errors for variable X1 and X2 for different sample size and different missing rate.

method of CCA to reduce the number of data and cannot be found close to the actual values of parameters. For this reason, if you want to use CCA which has very easy to apply, your data sets must have much sample size (N=100) and the highest proportion must 0,20 missing. All the missing data analysis methods can be used for the sample size is little and 5% and 10% mis sing rate while REG and MI give the closest value to the true regression coefficient in over 20% missing rate for sample of 25 units. For sample of 50 units, best methods are EM and MI. All the methods can be used for large sample size and less than 20% missing rate. Estimates of regression coefficients after performing MI, REG and EM closer to actual value than others method when the missing rate is over 40%.

The difference between the techniques grew for increasing missing rate and while the sample size increased the methods were similar to each other. The difference between estimates after performing missing data analysis method and estimates of complete (original) data sets was reduced.

CCA was the most affected method from sample size. The estimates after performing REG, EM and MI methods were very similar to each other and real value. Multiple imputation method as impute more than one value for each missing value should be preferred instead of single imputation method (mean imputation, regression imputation, EM imputation) as impute only one value for each missing value.

REFERENCES

1. Cox DR. Regression models and life tables. Journal of the Royal Statistical Society 1972; 34(2):187-220.
2. Little RJA. Regression with missing X's: a review. Journal of the American Statistical Association 1992;87(420):1227-37.
3. Allison PD. Multiple imputation for missing data: a cautionary tale. Sociological Methods and Research 2000;28(1):301-9.
4. Hosmer DW, Lemeshow S. Applied Survival Analysis: Regression Modelling of Time to Event Data. 1st ed. USA: John Wiley&Sons; 1999. p.271-99.
5. Rubin DB. Inference and missing data. Biometrika 1976;63(3):581-92.
6. Little RJ, Rubin DB. Statistical Analysis with Missing Data. 2nd ed. New York: John Wiley & Sons; 2002. p.1-340.
7. Enders CK. Applied Missing Data Analysis. 1st ed. New York: Guilford Press; 2010. p.1-347.
8. Schafer JL. Analysis of Incomplete Multivariate Data. 1st ed. London: Chapman&Hall; 1997. p.1-430.
9. Haitovsky Y. Missing data in regressionanalysis. Journal of the Royal Statistical Society Ser. B 1968;30(1):67-82.
10. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Ser B 1977;39(1):1-38.
11. Rubin DB. Multiple Imputation for Nonresponse in Surveys. 1st ed. New York: John Wiley&Sons; 1987. p.303.
12. Rubin DB. Multiple imputation after 18+ years (with discussion). Journal of the American Statistical Association 1996;91(434):473-89.
13. Schafer JL, Olsen MK. Multiple imputation for multivariate missing data problems: a data analyst's perspective. Multivariate Behavioural Research 1998;33(1):545-71.
14. White IR, Royston P. Imputing missing covariate values for the Cox model. Stat Med 2009;28(15):1982-98.
15. Marshall A, Altman DG, Holder RL. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. BMC Med Res Methodol 2010;10:112. doi: 10.1186/1471-2288-10-112.