Pattern-Mixture model of the Cox proportional hazards model with missing binary covariates

Tae-Mi Youk[1], and Juwon Song[1,2]
[1] Korea University, Seoul, Korea
[2] Corresponding author: Juwon Song, e-mail: jwsong@korea.ac.kr

Abstracts

When fitting the Cox proportional hazards model with missing covariates, it is inefficient to exclude observations with missing values in the analysis. Furthermore, if the missing-data mechanism is not Missing Completely At Random (MCAR), it may lead to biased parameter estimation. Many approaches have been suggested to handle the Cox proportional hazards model when covariates are sometimes missing, but they are based on the selection model. This paper suggests an approach to handle Cox proportional hazards model with missing covariates by using the pattern-mixture model (Little, 1993). The pattern-mixture model is expressed by the joint distribution of survival time and missing-data mechanism. In the pattern-mixture model, many models can be considered by setting up various restrictions, and different results under various restrictions indicate sensitivity of the model due to missing covariates. A simulation study was conducted to show the sensitivity of parameter estimation under different restrictions in pattern-mixture model.

Keywords: missing-data mechanism, sensitivity analysis, survival time, restrictions

1. Introduction

If a $(n \times v)$ data matrix $Z = \{z_{ij}\}$ is fully observed, v variables $Z = (Z_1, \ldots, Z_v)$ can be handled with standard techniques. However, observations with missing values can cause a biased estimate. In this case, it is normally used to remove units included missing values, called complete-case (CC) analysis. This method is simple, but inefficient because of the loss of information. Furthermore, in order to obtain an unbiased estimator, the strong assumption is required that the missing-data mechanism is Missing Completely At Random (MCAR) (Little and Rubin, 2002).

Missing value problems in the Cox proportional hazards model are related to censored failure time or covariates with missing. Cox (1972; 1975) proposed a parameter estimation to maximize partial likelihood function for the incomplete survival data by censoring without missing covariates. When estimating the parameter in the model with missing covariates, the missing-data mechanism should be considered as well as censoring.

That model expresses a joint distribution of the Cox proportional hazards model and the distribution related missing-data mechanism, most of the study assumed that this joint distribution of variables and the missing-data mechanism is based on the selection model (Hogan and Laird, 1997). Lin and Ying (1993) suggested an estimation to maximize the approximate partial likelihood function under MCAR, Chen and Little (1999) established a non-parametric maximum likelihood estimator (NPMLE) when the missing-data mechanism is Missing At Random (MAR). Herring, Ibrahim, and Lipsitz (2004) specified a parametric distribution for covariates and the missing-data mechanism, proposed a Monte carlo EM algorithm to compute estimates under the missing-data mechanism is Not Missing At Random (NMAR).

In this paper, we consider a pattern-mixture model (Little, 1993) for the joint distribution of variables in Cox proportional hazards model and missing-data mechanism and calculate parameters. In addition, we examine the effect of missing

covariates on the parameter estimation by various restrictions.
Section 2 describes the pattern-mixture model and several restrictions. Section 3 discusses the pattern-mixture model of the Cox proportional hazards model with missing binary covariates. Section 4 presents simulation results examining the sensitivity of parameter estimation under different restrictions in pattern-mixture model. Section 5 describes discussion and other issues.

2. Pattern-Mixture model

Incomplete data analysis is based on the joint distribution of missing indicator matrix $M = \{m_{ij}\}$ and a data matrix Z, there are two methods to model this distribution, the selection model and patter-mixture model (Little and Rubin, 2002). The selection model specifies that the joint distribution of M and Z satisfies

$$P(Z, M|\theta, \psi) = P(Z|\theta)P(M|Z, \psi)$$

where (θ, ψ) are, respectively, parameters for the marginal distribution of Z and the conditional distribution of M given Z. In this case, assume that the conditional distribution of M given Z is under MCAR, MAR, or NMAR, if the parameter θ and ψ are distinct and the missing-data mechanism is MCAR or MAR, then the conditional distribution of M given Z is ignorable (called ignorable missing data mechanism).
On the order hand, in the pattern-mixture model, the joint distribution of M and Z expresses

$$P(Z, M|\phi, \pi) = P(M|\pi)P(Z|M, \phi)$$

where (π, ϕ) are unknown parameters for marginal distribution of M and the conditional distribution of Z given M. The pattern-mixture model facilitates flexuous analysis under not MCAR. The pattern is a shape of occurred missing, is decided by unit missing indicator $(m_{i1}, \dots, m_{iv})$. Units holding same missing indicator belong to same pattern, completely observed pattern P_0 and incomplete pattern $P_1, \dots, P_{2^v-1}$ are theoretically possible. In general, $T + 1$ patterns exist for $T \leq 2^v - 1$.
Let the number of units in pattern P_t $(0 \leq t \leq T)$ is n_t $\left(\sum_{t=0}^{T} n_t = n\right)$ and $z_{obs,i}^{(t)}$ are observed variables and $z_{mis,i}^{(t)}$ are unobserved variables for i^{th} $(i = 1, \dots, n)$ unit. In addition, r_i indicates whether the pattern of i^{th} unit is r^{th} pattern, then the probability is $p(r_i = r) = \pi_r$. In this case, r_i follows a multinomial distribution and the distribution of z_i given r_i divide an observed z_i and a conditional part of unobserved z_i given observation, that is,

$$P\left(z_i|r_i = r, \phi^{(r)}\right) = P\left(z_{obs,i}^{(r)}|r_i = r, \phi_{obs,i}^{(r)}\right) P\left(z_{mis,i}^{(r)}|r_i = r, z_{obs,i}, \phi_{mis,i}^{(r)}\right),$$

where $\phi^{(r)}$ is a interest parameter for the pattern P_r. $\phi_{obs,i}^{(r)}$ and $\phi_{mis,i}^{(r)}$ are the function of $\phi^{(r)}$, not-missing and missing respectively. The observed likelihood function with missing specifies

$$L(\pi, \phi|Z\ M) = L(\pi, \phi|Z_{obs}, M) = \prod_{r=0}^{T} \left\{ \pi_r^{n_r} \prod_{i \in P_r} P\left(z_{obs,i}^{(r)}|r_i = r, \phi_{obs,r}^{r}\right) \right\}$$

and we estimate a parameter to maximize the function.
The pattern-mixture model is more honest than the selection model in that parameters are regarded by differently; it is split between estimable and inestimable parameters whether to be observed. But there is a problem with identifiability about inestimable parameters. The way to solve this problem is to assume that inestimable parameters in the incomplete pattern are equal to functions of parameters in the observed pattern applying a species of restriction. The pattern-mixture model has a point to do not complicate fitting the model based on the data even if restrictions are included (Wang and Daniels, 2010).
The commonly considered restriction is Complete-Case Missing-Variable (CCMV) that assumes identifiable parameters for the incomplete patterns are the same as a parameter for the fully observed pattern P_0, that is,

$$\phi_{mis,r}^{(r)} = \phi_{mis,r}^{(0)}.$$

In this respect CCMV restriction in the pattern-mixture model relate with the selection model under MCAR. It is expedient to use CCMV restriction when most belong to the pattern P_0 and only a small percentage of units are in incomplete patterns. Furthermore, it is extensible partially through a non-monotone form (Thijs, Molenberghs, Michiels, Verbeke and Curran, 2002).

Alternative to CCMV restriction is Available-Case Missing-Variable (ACMV). This restriction specifies a conditional distribution of missing data given observation is equal to responded whole pattern. In detail, s $(s \leq t)$ subsets are formed by grouping patterns and then the assumption is identification of parameters for each subset using only responded units. Let $r^{(r)}$ be a parameter indexing the distribution of z_i for the pattern P_r in the saturated pattern-mixture model. Then $r^{(r)}$ is equated to the set of patterns S if

$$r^{(r)} = \sum_{s \in S} \pi_s r^{(s)} / \sum_{s \in S} \pi_s$$

where π_s is a probability that unit pattern is P_s.

The restriction such as CCMV or ACMV in the pattern-mixture model is related to sensitivity showing uncertainty in estimate. A part of restrictions correspond to the selection model under MCAR, MAR, or NMAR (Molenberghs, Michiels, Kenward and Diggle, 1998; Wang and Daniels, 2010).

3. Pattern-Mixture model with missing binary covariates

In the survival data is comprised of v covariates, survival time t, and censoring indicator δ, if all covariates for i^{th} unit are fully observed then it has the missing indicator $m_i = (0, 0, \dots, 0)$, otherwise the missing indicator is a combination of zero and one. The pattern is decided by a form of missing indicator, P_0 is a set of units which it have the missing indicator $(0, 0, \dots, 0)$ and another case are included in one of $P_1, \dots, P_T$.

When the missing data mechanism is MCAR, it is possible to analyze the pattern-mixture model with CCMV restriction because the distribution for pattern P_0 is equal to the distribution of incomplete data patterns. In detail, each incomplete data pattern estimates parameters of a model to impute missing covariates based on the pattern P_0, and then predictive value is used in place of missing value. When two and more missing covariates exist in same incomplete pattern, the joint distribution of covariates specifies the product of one-dimension conditional distributions (Herring and Ibrahim, 2001). If missing covariate is continuous, a conditional distribution of unobserved covariate given observed value able to set the model such as linear regression model. Likewise, predictive probability is calculated by logistic regression model when missing covariate is binary.

The pattern-mixture model is applied to Cox proportional hazard model with missing binary covariate, in this paper, that incomplete data is imputed by predictive probability under various restrictions; the stochastic imputation is possible to include error term. After imputation, it is able to fit the Cox proportional hazards model with non-missing covariates.

In the monotone pattern data, the selection model under MAR represents that cause of missing depends on observed data set, which it is related to ACMV restriction in that the conditional distribution of unobserved data given observed data is equal to a subset of whole pattern (Molenberghs, Michiels, Kenward and Diggle, 1998). In addition, the pattern-mixture model with ACMV restriction has a different effect by composition of subset.

4. Simulation study

A simulation study was conducted as follows to show the sensitivity of parameter estimation under different restrictions in pattern-mixture model. Survival time t_i was generated from exponential distribution with parameter $\lambda_i = \exp(z'_i \beta)$ and non-

censoring distribution was considered both randomly 30% and non-censoring distribution. We considered two binary covariates; first covariate z_1 was generated from bernoulli with 0.5 and an odds ratio of covariates was 9. The sample size is 200. We first consider monotone pattern data set that z_1 is fully observed and z_2 has missing units. When the missing-data mechanism is MAR, long follow-up time is deleted by 50%. Second, we determined missing value under MAR assumption by stochastic selection step,

$$\text{logit}[p(m_1 = 1)|t, \delta, z_1, z_2] = -3t/2$$
$$\text{logit}[p(m_2 = 1)|t, \delta, z_1, m_1, z_2] = -t - (1 - m_1)z_1.$$

For comparison, we fitted the Cox proportional hazard model without missing and also calculated parameter estimation for Complete-Case (CC) analysis. The first simulation, monotone pattern data set, results are presented in Table 1. When the missing-data mechanism was MCAR, estimated regression coefficients of CC and CCMV were similar with true value; however, standard error of CC estimate was greater than CCMV or not-missing case. Furthermore, parameters in the pattern-mixture model were underestimated when the absolute value of true parameter is more than two. Under MAR, CC analysis offered seriously biased estimates in almost cases while results of CCMV method were closer to the true value than CC. These results are from degree of using information, CC method is less accurate because it estimates parameters using only short follow-up time data. On the other hand, the pattern-mixture model with restriction utilizes all available data and then remedies CC's weaknesses. In addition, in common with MCAR, if the absolute value of true parameter is increasing, parameters is underestimated because the logic model for imputing missing covariate is affected by a difference from real proportional hazard model. When the censoring rate is 0.3, results was similar with non-censoring case but was lower the accuracy by loss of information.

Table 1. Parameter estimates in the monotone pattern data set that z_1 is fully observed and z_2 has missing units without censoring

		Not missing		CC		CCMV	
	TP	β_1	β_2	β_1	β_2	β_1	β_2
MCAR	(0,0)	-0.002[1)]	-0.002	0.002	-0.012	0.004	-0.010
		(0.174)[2)]	(0.160)	(0.252)	(0.230)	(0.175)	(0.162)
	(1,0)	1.015	0.001	1.019	-0.004	1.013	0.011
		(0.191)	(0.160)	(0.277)	(0.231)	(0.192)	(0.162)
	(1,1)	1.006	1.007	1.025	1.020	0.971	1.002
		(0.192)	(0.173)	(0.279)	(0.251)	(0.195)	(0.178)
	(1,−1)	1.005	-1.013	1.014	-1.032	0.928	-0.959
		(0.187)	(0.172)	(0.270)	(0.249)	(0.182)	(0.170)
	(2,−2)	2.005	-2.006	2.016	-2.020	1.725	-1.755
		(0.218)	(0.203)	(0.315)	(0.295)	(0.202)	(0.191)
MAR	(0,0)	-0.002	-0.002	-0.003	0.002	0.032	-0.014
		(0.174)	(0.160)	(0.251)	(0.229)	(0.172)	(0.169)
	(1,0)	1.015	0.001	0.165	-0.015	1.020	-0.091
		(0.191)	(0.160)	(0.298)	(0.223)	(0.187)	(0.167)
	(1,1)	1.006	1.007	0.116	0.207	0.986	0.845
		(0.192)	(0.173)	(0.335)	(0.244)	(0.199)	(0.188)
	(1,−1)	1.005	-1.013	0.249	-0.254	0.759	-0.961
		(0.187)	(0.172)	(0.265)	(0.227)	(0.174)	(0.173)
	(2,−2)	2.005	-2.006	0.756	-0.769	1.524	-2.147
		(0.218)	(0.203)	(0.287)	(0.239)	(0.192)	(0.205)

NOTE: 1000 iterations; TP: true parameter value; CC: complete-case analysis; CCMV: complete-case missing-variable restriction; [1)] : mean of parameter estimate; [2)] : mean of standard error for estimate

Table 2. Parameter estimates under MAR assumption by stochastic selection step without censoring

					Pattern-mixture					
	Not missing		CC		CCMV		ACMV1		ACMV2	
TP	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
(0,0)	-0.007[1)]	0.006	0.086	0.006	0.039	-0.010	0.026	-0.011	0.000	0.000
	(0.160)[2)]	(0.175)	(0.213)	(0.236)	(0.161)	(0.177)	(0.161)	(0.176)	(0.161)	(0.176)
(0,1)	-0.002	1.015	0.073	1.228	0.044	1.027	0.028	1.022	-0.001	1.035
	(0.160)	(0.191)	(0.244)	(0.287)	(0.163)	(0.195)	(0.162)	(0.194)	(0.161)	(0.193)
(0,2)	-0.008	2.038	0.046	2.520	0.002	2.079	-0.008	2.065	-0.032	2.079
	(0.160)	(0.240)	(0.268)	(0.414)	(0.164)	(0.247)	(0.163)	(0.245)	(0.162)	(0.244)
(1,1)	0.999	1.023	1.269	1.248	1.024	1.026	1.012	1.028	0.979	1.040
	(0.173)	(0.192)	(0.288)	(0.303)	(0.178)	(0.198)	(0.177)	(0.197)	(0.175)	(0.196)
(2,2)	2.018	2.019	2.339	2.459	1.787	1.903	1.781	1.910	1.744	1.925
	(0.215)	(0.234)	(0.423)	(0.424)	(0.213)	(0.245)	(0.212)	(0.244)	(0.209)	(0.244)
(−1,1)	-1.010	1.009	-1.082	1.172	-0.891	0.922	-0.911	0.919	-0.940	0.933
	(0.172)	(0.187)	(0.249)	(0.267)	(0.173)	(0.188)	(0.173)	(0.187)	(0.172)	(0.186)
(−2,2)	-2.008	2.013	-2.283	2.361	-1.549	1.590	-1.573	1.591	-1.590	1.592
	(0.204)	(0.218)	(0.325)	(0.340)	(0.190)	(0.205)	(0.190)	(0.204)	(0.189)	(0.203)

NOTE: 1000 iterations; TP: true parameter value; CC: complete-case analysis; CCMV: complete-case missing-variable restriction; ACMV: available-case missing-variable restriction; [1)] : mean of parameter estimate; [2)] : mean of standard error for estimate

The second simulation results are shown in Table 2. The data is divided into four patterns, P_0, P_1, P_2, or P_3, it is that missing indicator is (0,0), (0,1), (1,0), and (1,1) respectively. A percentage that only one of two covariates is missing is 15.4~22.3% and a percentage that both two covariates are missing is 9.9~19.9%.

ACMV1 restriction is a method which, first, we impute missing values in P_0 and P_1 by CCMV restriction and then a pooled pattern P_{012} is created to fit the logic model in pattern P_3. The pattern-mixture model with ACMV2 sequentially apply P_{01} and P_{012} using CCMV restrictions. When true parameter has at least one 0, estimates in the pattern-mixture model considering restrictions was closer to true parameters than CC analysis. These results are caused by loss of information reflected only part of observed data even though variable is not significant in the model. ACMV1 and ACMV2 that use different data sets have shown similar results.

Consequentially, when the missing-data mechanism is MCAR, CC analysis and pattern-mixture method offered very identical performance, but CC method failed to estimate parameters in almost cases under MAR. In addition, the pattern-mixture model was more stable than CC estimation. However, single imputation of missing value is able to undermine the credibility of results when the absolute value of true parameter increases and if rate of missing is high, it could lead to a wrong conclusion.

5. Conclusions

In order to solve a problem with identifiability about inestimable parameters in the pattern-mixture model, we could apply a species of restriction to the model such as CCMV or ACMV. The pattern-mixture model is known as being very sensitive to misspecification of the model (Demirtas, 2005). A simulation study was conducted to show the sensitivity of parameter estimation under different restrictions in pattern-mixture model. Furthermore, fitting the pattern-mixture model of the Cox proportional hazards model with missing binary covariates provided an estimate in the middle between not-missing and Complete-Case analysis. This paper showed that applying the pattern-mixture model would be better than removing missing response under

MCAR or some of MAR. To the extent missing values are being imputed under particular restrictions, it is also possible to expend into using Monte Carlo Markov Chain (MCMC). In addition, this paper shows the sensitivity of parameter estimation under different restrictions in pattern-mixture model and tries to make an interpretation as compared with the selection model because incomplete data is generally analyzed under MCAR or MAR assumptions.

Ultimately, the pattern-mixture model is a technique applied when the missing-data mechanism is NMAR for considering in the shape of missing. This paper provides a framework for further study that is the sensitivity analysis of parameter estimation in pattern-mixture model with various restrictions under NMAR.

References

Chen, H. Y., and Little, R. J. A. (1999)"Proportional hazards regression with missing covariates,"*Journal of the American Statistical Association*, 94, 896-908.

Cox, D. R. (1972)"Regression models and life-tables (with discussion),"*Journal of the Royal Statistical Society*, Ser. B, 34, 187-220.

Cox, D. R. (1975)"Partial likelihood,"*Biometrika*, 62,269-279.

Demirtas, H. (2005)"Multiple imputation under Bayesianly smoothed pattern-mixture models for non-ignorable drop-out,"*Statistics in Medicine*, 24, 2345-2363.

Herring, A. H., and Ibrahim, J. G. (2001)"Likelihood-based methods for missing covariates in the Cox proportional hazards model,"*Journal of the American Statistical Association*, 96, 292-302.

Herring, A. H., Ibrahim, J. G., and Lipsitz, S. R. (2004)"Non-ignorable missing covariate data in survival analysis: a case-study of an International Breast Cancer Study Group trial,"*Journal of the Royal Statistical Society*, 53, 293-310.

Hogan, J. W., and Laird, N. M. (1997)"Model-based approaches to analysing incomplete longitudinal and failure time data," *Statistics in Medicine*, 16, 259-272.

Kalbfleisch, J. D., and Prentice, R. L. (1980) *The Statistical Analysis of Failure Time Data*, New York: Wiley.

Lin, D. Y., and Ying, Z. (1993)"Cox regression with incomplete covariate measurements," *Journal of the American Statistical Association*, 88, 1341-1349.

Little, R. J. A. (1993)"Pattern-Mixture models for multivariate incomplete data," " *Journal of the American Statistical Association*, 88, 125-134.

Little, R. J. A., and Rubin, D. B. (2002) *Statistical Analysis With Missing Data*, New York: Wiley.

Molenberghs, G., Michiels, B., Kenward, M. G., and Diggle, P. J. (1998)"Monotone missing data and pattern-mixture models,"*Statistica Neerlandica*, 52, 153-161.

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., and Curran, D. (2002)"Strategies to fit pattern-mixture models,"*Biostatistics*, 3, 245-265.

Wang, C., and Daniels, M. J. (2011)"A note on MAR, identifying restrictions, and sensitivity analysis in Pattern-Mixture models with and without covariates for incomplete data,"*Biometrics*, 67, 810-8.