Proceeding of ICSA 2019, p: 83-95
ISBN 978-979-19256-3-1 (PDF)

Cox Piecewise Constant Hazard Model with Bayesian Method

Amanda Putri Tiyas Pratiwi[1], Sarini Abdullah[2*], Ida Fithriani[3]

[1,2,3]*Mathematics, Universitas Indonesia, Depok, 16424, Indonesia*

[*)]Corresponding author: sarini@sci.ui.ac.id

Abstract

Cox PH model is one of the survival models that is widely used for analyzing time-to-event data. Cox PH model consists of two main components, the baseline hazard consisting of time-dependent component; and the exponential function accomodating explanatory variables. The baseline hazard is not estimated in the Cox PH model, thus not accommodating the need for hazard rate estimation. Therefore, in this paper we discuss the estimation of baseline hazard through piecewise constant hazard using Bayesian method. Gamma distribution is assumed for the piecewise constant baseline hazard, and normal distribution is assumed for the regression coefficient. Sampling from the posterior is conducted using Markov chain Monte Carlo through Gibbs sampling. Echocardiogram data containing 106 observations and 6 explanatory variables were used in analysis. The result showed that the baseline hazard functions were estimated and each of parameters in the model is converged as shown by the trace plot and posterior density plot.

***Keywords*:** Bayesian method, Cox regression, Gibbs sampling, Markov chain Monte Carlo, survival analysis.

Introduction

Survival analysis is a method to analyze time-to-event data. Normally, time-to-event data associate with additional information (explanatory variables). In analyzing time-to-event data, probability of survive or failure risk of subject to experience the event of interest would be the focus of study. Time-to-event data modelling is needed in order to know the subject's survival probability or failure risk with corresponding to the explanatory variables.

One of the most widely used model in analyzing time-to-event data correspond to the explanatory variables is Cox model (Guo and Zeng, 2013). By using Cox model, hazard (failure risk) corresponding to the explanatory variables can be estimated. Cox model has two main components, baseline hazard and exponential function that includes regression coefficient. In this model, the baseline hazard does not have any assumption (Klein and Moeschberger, 2003). Therefore, in order to determine subject's hazard specifically, this baseline hazard should be specified.

In this paper, the baseline hazard is assumed to be piecewise constant. By assuming piecewise constant for the baseline hazard, we have a model called piecewise constant hazard model. By using piecewise constant hazard model, hazard of each subject can be determined specifically. Afterward, Bayesian method is employed to estimate parameters in piecewise constant hazard model in which sample of parameters will be obtained using Markov chain Monte Carlo through Gibbs sampling. The construction of piecewise constant hazard model will be explained and the model will be applied to real data. Echocardiogram data from Universitify of California, Irvine (UCI) consist of 106 observation and 6 explanatory variables are used in analysis. The six variables consist of age, pericardial-effusion, fractional-shortening, EPSS, LVDD and wall-motion-index. Afterward, convergence checking of all parameters in the model is conducted. Parameters convergence can be seen through trace plot and density plot of model diagnostics. The result shows that all of parameters in the model is converge and

therefore the estimated value of parameters will be used to determine hazard of a subject specifically.

David Cox made a breaktrough with his research titled Regression Models and Life-Tables in 1972. This research focus on modelling time-to-event data with considering the explanatory variables in order to estimate hazard. This research considered to use censored failure times and assumed that each individual has values on explanatory variables. Cox model defined hazard as a product of an unknown function of time and exponential function that includes regression coefficient. The research also include an explanation about conditional likelihood that leading to inferences about the regression coefficients. This procedure of conditional likelihood does not need the specification of function of time in Cox model, made the model flexible and easy to use.

However, the drawback of this method is that Cox PH model can only be used to estimate hazard ratio. Without knowing about the unkown function of time in this model, spesific hazard of each individual can not be estimated. For this purpose, the unknown function of time should be assumed to have a spesific form or function. This purpose can lead to other research that can be done by making assumption regarding the unknown function of time.

Materials

Piecewise constant hazard model is obtained from the piecewise constant assumption for the baseline hazard in Cox PH model. Cox PH model consist of two components, baseline hazard which is just based on time and exponential function that includes regression coefficient in which this expoential function just based on explanatory variables. In Cox PH model, the baseline hazard does not have any assumption. Therefore, we can not determine a subject's hazard because we have unknown component in this model.

Cox PH model is defined as follow. Assumed T is survival-time and each subject has p explanatory variables with $\boldsymbol{Z}^t = [Z_1, Z_2, \ldots., Z_p]$. Then, hazard of a subject based on their explanatory variables defined as :

$$h(t|\boldsymbol{Z}) = h_0(t).e^{(\boldsymbol{\beta}^t \boldsymbol{Z})} \quad (1)$$

(Cox,1972)

In Cox model, $h_0(t)$ defined as baseline hazard in which when the explanatory varaiables assumed have values of zero or when the explanatory variables are not considered in the model. Then, $\boldsymbol{\beta}^t = [\beta_1, \beta_2, \ldots., \beta_p]$ is a vector of regression coefficient, expressing the effects of explanatory variables and $\boldsymbol{Z}^t = [Z_1, Z_2, \ldots., Z_p]$ is a vector of explanatory variables corresponded to the subject. By Cox model, hazard of a subject based on their explanatory variables defined as the product of baseline hazard and an exponential function which contain regression coeffients. In order to dertermine hazard, both the baseline hazard and regression coefficient in this model need to be estimated. In this paper, parameter estimation is done by Bayesian method through Markov chain Monte Carlo and Gibbs sampling.

In Bayesian method, parameters of interest is treated as random variables. Bayesian method use the information of observed data, in the form of likelihood function and the information of historical data or historical information about parameter of interest in the form of prior distribution. Then, the parameters of interest can be obtained from posterior distribution.

Let θ denotes the parameter of interest and D is observed data. Assumed that θ has prior distribution denoted by $\pi(\theta)$. Estimated parameter in Bayesian method will be obtained from posterior distribution, denoted by $\pi(\theta|D)$ (Ibrahim, Chen and Sinha, 2011). Posterior distribution is given by:

$$\pi(\theta|D) = \frac{\pi(D|\,\theta).\pi(\theta)}{\int_\theta \pi(D|\,\theta).\pi(\theta)\,d\theta} \quad (2)$$

where $\pi(D|\theta)$ is likelihood from observed data. The denominator of posterior distribution called normalization constant. If the value is known, it is just a constant that made posterior distribution has pdf property. Therefore, posterior distribution can be written in proportional form as the product of likelihood and prior (Ibrahim, Chen and Sinha,2011) as follow :

$$\pi(\theta|D) \propto \pi(D|\theta).\pi(\theta) \quad (3)$$

Posterior distribution can take one of this two forms, closed form and non-closed form. The estimated parameter can be obtained directly from closed form posterior while in the non-closed form posterior, samples of parameter will be obtained with through Markov chain Monte Carlo method (Hoff,2009). In this paper, Markov chain Monte Carlo method with Gibbs sampling alghorithm will be used as the obtained posterior distribution take on a non-closed form. Therefore, parameter estimation of piecewise constant hazard model is conducted with Markov chain Monte Carlo and Gibbs sampling. In piecewise constant hazard model, there are two parameters that need to be estimated, baseline hazard in the form of piecewise constant and regression coefficient.

In piecewise constant hazard model, time axis is divided into several interval in which every interval has its own baseline hazard in the form of a constant. Then, subjects in study is divided based on their survival time into one of these intervals. Therefore, each individu has baseline hazard based on the interval they belong to. In piecewise constant hazard model, hazard of a subject is defined as the product of a constant and exponential function that includes regression coefficient. In this paper, it is assumed that explanatory variables are time-independent and each parameter in the model is disjoint.

Assumed that random variable T denotes survival time and each subject has p explanatory variables denoted by $\boldsymbol{Z}^t = [Z_1, Z_2, \ldots, Z_p]$. Assumed that baseline hazard in Cox

model is constant, which denoted by $h_0(t) = \lambda$. Then, assumed *D* is observed data with *n* subjects and *p* explanatory variables. Piecewise constant hazard model defined as :

$$h_{ij}(t|\boldsymbol{Z}) = \lambda_j . e^{(\boldsymbol{\beta^t Z})} \qquad i = 1,2,\dots,n, \quad j = 1,2,\dots.,J. \tag{4}$$

(Ibrahim, Chen and Sinha,2011)

The time axis is made into finite partition where $0 < S_1 < S_2 < \cdots < S_J$ with $S_J > y_i$ for all $i = 1,2,\dots,n$ where y_i is the survival time of the subject. In Equation (4), the survival time lies in between interval S_{j-1} and interval S_j or denoted as $t \in (S_{j-1}, S_j]$. The notation $i = 1,2,\dots,n$ denotes subjects in the study and $j = 1,2,\dots,J$ denotes the interval. In order to determine the hazard, both constant and regression coefficient need to be estimated. Once the constant and regression coefficient are estimated, the hazard of each subject can be determined. In order to do this, we need the component likelihood function and prior distribution in order to get posterior distribution.

In this paper, the assumption of censored type I of time-to-event data is used to construct likelihood function based on piecewise constant hazard model. When the subject's survival time is known, the value of y_i will be t_i and when the subject is censored the value of y_i will be c_i. Mathematically, it can be written as :

$$y = \min(t, c) \tag{5}$$

In order to construct likelihood for this model, the pdf and survival function of this model is needed. For piecewise constant hazard model, survival function defined as :

$$S(t|\boldsymbol{Z}) = e^{-[\lambda_j(y_i - s_{j-1}) + \sum_{g=1}^{j-1} \lambda_g (S_g - s_{g-1})].e^{(\boldsymbol{\beta^t Z})}} \tag{6}$$

and the pdf has the form:

$$f(t|\boldsymbol{Z}) = \left[\lambda_j . e^{(\boldsymbol{\beta^t Z})}\right].\left[e^{-[\lambda_j(y_i - s_{j-1}) + \sum_{g=1}^{j-1} \lambda_g (S_g - s_{g-1})].e^{(\boldsymbol{\beta^t Z})}}\right] \tag{7}$$

Then, defined random variable v that denotes status of the subject and random variable δ_{ij} denotes status of the subject per interval, defined as follow :

$$v = \begin{cases} 1, \text{if the subject survival time is observed } (t \leq c) \\ 0, \text{if the subject is censored } (t > c) \end{cases} \quad (8)$$

$$\delta_{ij} = \begin{cases} 1, \text{if the subject experience the event or censored in } j-th \text{ interval} \\ 0, \text{others} \end{cases} \quad (9)$$

By using these forms of pdf, survival function and assumption that observed data has censored Type I observation, likelihood function for piecewise constant hazard model defined as follow :

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D) = \prod_{i=1}^{n} \prod_{j=1}^{J} \frac{\left[h_0(y_i).\, e^{(\boldsymbol{\beta}^t \boldsymbol{Z})}\right]^{\delta_{ij} v_i}}{\left[e^{-\delta_{ij}[\lambda_j (y_i - s_{j-1}) + \sum_{g=1}^{j-1} \lambda_g (S_g - s_{g-1})].e^{(\boldsymbol{\beta}^t \boldsymbol{Z})}}\right]} \quad (10)$$

After having the form of likelihood function for piecewise constant hazard model, the prior distribution should be specified for each parameter in the model. If θ denotes parameter, then in piecewise constant hazard model we have $\boldsymbol{\theta}^t = (\boldsymbol{\beta}, \boldsymbol{\lambda})$ where $\boldsymbol{\beta}^t = [\beta_1, \beta_2, \ldots., \beta_p]$ and $\boldsymbol{\lambda}^t = [\lambda_1, \lambda_2, \ldots., \lambda_J]$. In this paper, it is assumed for each of β_i for $i = 1,2, \ldots, p$ will be following normal distribustion with $\mu = 0$ and $\sigma^2 = 10^6$. Assumed that for each β is disjoint, prior distribution for parameter **β** is :

$$\pi(\boldsymbol{\beta}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^p \exp\left\{-\frac{\left(\sum_{i=1}^{p} \beta_i{}^2 - 2\mu \sum_{i=1}^{p} \beta_i + \sum_{i=1}^{p} (\mu)^2\right.}{(2\sigma^2)}\right\}, (-\infty < \beta_i < \infty), \quad i=1,2,...p \quad (11)$$

Subsequently, for each $j = 1, 2,3, \ldots., J$, it is assumed that each of parameter λ will be following gamma distribution with $\alpha = 10^{-4}$ and $\gamma = 10^{-4}$. Assumed that for each λ is disjoint, prior distribution for parameter **λ** is :

$$\pi(\boldsymbol{\lambda}) = \frac{(\gamma^{\alpha})^p}{(\Gamma(\alpha))^p} \prod_{j=1}^{J} (\lambda_j)^{\alpha-1} e^{-\gamma(\sum_{j=1}^{J} \lambda_j)}, \quad \alpha > 0, \gamma > 0, 0 < \lambda_j < \infty, j = 1,2,3, \ldots, J \quad (12)$$

After having the likelihood function and prior distribution for all parameters, by using the proportional form, posterior distribution for piecewise constant hazard model defined as :

$$\pi(\boldsymbol{\beta},\boldsymbol{\lambda}\,|\mathrm{D}) \propto \prod_{i=1}^{n}\prod_{j=1}^{J}\lambda_j . e^{(\boldsymbol{\beta}^t \boldsymbol{Z})}]^{\delta_{ij}v_i} . [e^{-\delta_{ij}[\lambda_j(y_i - s_{j-1}) + \sum_{g=1}^{j-1}\lambda_j(S_g - s_{g-1})].e^{(\boldsymbol{\beta}^t \boldsymbol{Z})}}] .$$
$$(\frac{1}{\sigma\sqrt{2\pi}})^p \exp\{-\frac{(\sum_{i=1}^{p}\beta_i{}^2 - 2\mu\sum_{i=1}^{p}\beta_i + \sum_{i=1}^{p}(\mu)^2}{(2\sigma^2)}\} . \frac{(\gamma^\alpha)^p}{(\Gamma(\alpha))^p}\prod_{j=1}^{J}(\lambda_j)^{\alpha-1} e^{-\gamma(\sum_{j=1}^{J}\lambda_j)} \quad (13)$$

(Ibrahim, Chen and Sinha, 2011)

The posterior distribution above has a non-closed form. Therefore, estimated parameter can not be obtained directly and will be conducted by Markov chain Monte Carlo through Gibbs sampling. In Markov chain Monte Carlo, the samples of parameter are drawn using iteration which have limiting distribution to posterior distribution of interest. Model diagnostics using trace plot and density plot can show the convergence of the parameters of interest (Hoff,2009).

Method of Analysis

In this paper, piecewise constant hazard model is applied to echocardiogram data (UCI,1989). The data consists of 106 observations and 6 explanatory variable which are age, pericardial-effusion, fractional-shortening, EPSS, LVDD and wall-motion index . The subjects in the study are patients who had heart attack in the past. The event of interest in this study is the death of the patient because of the heart attack. The start of the study is from the last time the patient had heart attack in the past.

The procedure of data analysis is done as follow.

1. Define all the variables for the model based on the data.
2. Define model equation based on the parameters for echocardiogram data.
3. Determine number of iterations. In this study, the total number of iteration is 300.000, in order to get convergence for all parameters
4. Determine number of intervals for the model. This study use eight interval for the piecewise constant baseline hazard.
5. Determine prior distributions for each parameter in the model. Hyperparameter for prior also determined in this step. This study assume that prior distribution for

regression coefficients follow Normal$(0,10^6)$ and for the piecewise constant baseline hazard will follow Gamma$(10^{-4}, 10^{-4})$.

6. After all of the components are set, run the procedure.
7. Do convergence check for all parameters in the model by using trace plot and density plot.
8. If all the parameter hasn't converged, add number of iterations and re-run the procedure.
9. Procedure will stop if each of parameters in the model is converged.

After the procedure is done, the last step to do is to check the estimated value for the baseline hazards and regression coefficients in the model. After all the estimated parameter is obtained, the hazard of a specific subject can be determined.

Results and Discussion

Before the estimated parameter can be used in the model, parameter convergence should be checked first. In this paper, trace plot and density plot will be used as convergence diagnostic. For parameter β, the estimated parameter has converged if the trace plot and density plot has form as follow :

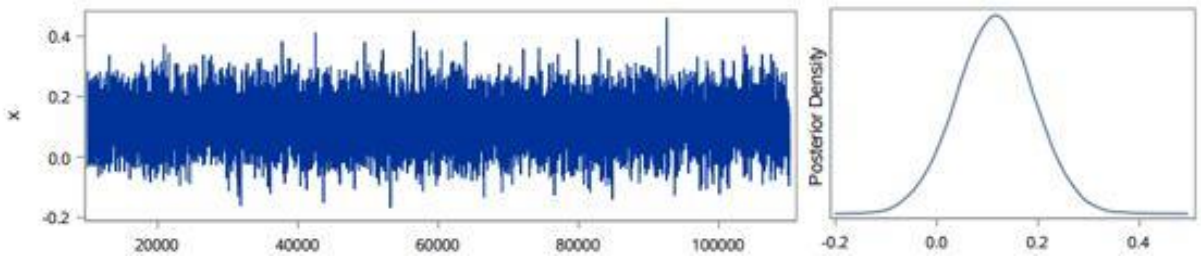

Figure1. Trace plot and density plot that shows the β parameter has converged

Subsequently for parameter λ, the trace plot and density plot should have form as follow :

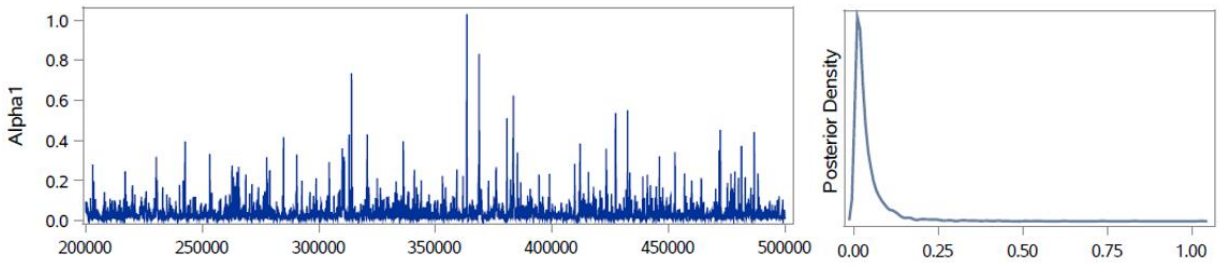

Figure2. Trace plot and density plot that shows parameter λ has converged

The result of analysis in this paper, all of the parameters has similar result of trace plot and density plot as the figures above. Therefore, all of the parameters have converged. In the output result, the obtained estimated parameter are as follow :

Table1. Estimated λ .

Interval	Estimated λ	
[0,16.5)	λ_1	0.039
[16.5,24.5)	λ_2	0.102
[24.5,265)	λ_3	0.393
[26.5,30)	λ_4	0.225
[30,33.5)	λ_5	0.319
[33.5,37)	λ_6	0.493
[37,48.5)	λ_7	0.301
[48.5, ∞)	λ_8	1.378

Table2. Estimated β.

Estimated β	
β_1	-0.015
β_2	-0.0001
β_3	-1.076
β_4	0.012
β_5	0.004
β_6	-0.0006

Because all of estimated parameters have converged, the estimated parameter can now be used to determine hazard. The hazard of each subject can now be determined. For instance, the subject of interest is subject number 60 who has information as follow :

Table 3. Information regarding subject number 60.

No.	Survival time	Censor	Age at heart attack	Pericardial-effusion	Fractional-shortening	EPSS	LVDD	Wall motion index
60	22	1	57	0	0,14	16,01	4,36	1,36

The subject died (censor=1), 22 months since the last time the subject had a heart attack (y=22) and based on Table 2, the survival time lies in the 2nd interval. Because this subject's survival time lies in interval 2, the value of baseline hazard will follow the value of λ_2. By using the infomation of explanatory variables, hazard of subject number 60 is :

$$h_{60,2}(t|\boldsymbol{Z}) = \lambda_2 . e^{(\boldsymbol{\beta^t Z})}$$

$$= (0.102).(e^{(-0.015)(57)+(0.017)(0)+\cdots+(-0.00006)(1.36)})$$

$$= (0.102).(0.45)$$

$$= 0.046$$

In the result above, the hazard of a subject can now determined specifically using piecewise constant hazard model based on their survival time and their corresponding explanatory variables. For instance above, the hazard of subject number 60 who died at the 22 months since the last time the subject had an heard attack is 0.046.

Cox model does not have any assumption regarding the baseline hazard. In order to determined hazard of each subject specifically, the baseline hazard need to be specified. Research regarding piecewise constant hazard moodel are still few. Research regarding piecewise constant hazard model is mostly applied to health and medicine data. However, the use of this model is not restricted only in that area. This model can be applied to other field such as economics or engineering. In this paper, model diagnostics for convergence of parameters had been discussed. This paper use trace plot and density plot to help determine the convergence of sampled parameter.

Conclusion and Future Research

In Cox model, specifying assumption for the baseline hazard can help to determine hazard per subject specifically. One of the approaches is to make a piecewise constant assumption for the baseline hazard in Cox model. By specifying this assumption, piecewise constant hazard model is obtained. In order to know the hazard of a subject, the constant (baseline hazard) and regression coefficient in Cox model need to be estimated. Bayesian method is employed to estimate parameters in the model with the help of Markov chain Monte Carlo through Gibbs sampling. The result showed that all of parameters is converge via diagnostics from trace plot and density plot for each parameters.

In this paper, assumption of censored type I is used. However, other types of time-to-event data can still be analyzed using piecewise constant hazard model. In this paper, the piecewise constant assumption is used for the baseline hazard. However, there are any other

varieties of assumption that can be made for baseline hazard assumption, such as gamma process, correlated gamma proces, beta process or Dirichlet process. Model diagnostics also not restricted by using only just trace plot or density plot. The other diagnostics also can be used to determine sampled parameters convergence such as autocorrelation plot, Geweke-Test or Gelman-Rubin test.

Acknowledgements

This research supported by the University of Indonesia with *PITTA B 2019* research grant scheme, with ID number NKB-0665/UN2.R3.1/HKP.05.00/2019. We thank to all reviewers for the improvement of this article.

References

Cox, D. R. (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187-202. [PDF]. September 30, 2018. http://remote-lib:ui:ac:id: 2093/stable/2985181.

Guo, S., and Zeng, D. (2013). An Overview of Semiparametric Models in Survival Analysis [PDF]. October 01, 2018.https://remote-lib:ui:ac:id:2053/science/ article/pii/S0378375813002619.

Hoff, P. D. (2009). First Course in Bayesian Statistical Methods (Springer texts in statistics). New York: Springer.

Ibrahim, J. G., Chen, M., and Sinha, D. (2011). Bayesian Survival Analysis. New York: Springer.

Klein, J. P., and Moeschberger, M. L. (2003). Survival Analysis Techniques for Censored and Truncated Data.Wiley.

Press,S. (2003). Subjective and objective Bayesian statistics: Principles, models, and applications. NJ: John Wiley and Sons.

SAS/STAT (2010). SAS/STAT(R) 9.2 User's Guide, Second Edition: Assessing Markov Chain Convergence. June 14, 2019. https://support:sas:com/documentation/ cdl/en/statug/63033/HTML/default/viewer:htm#statug_introbayes_sect008:htm.

UCI, (1989, February 28). Retrieved March 1, 2019, from https://archive.ics.uci.edu/ml/datasets/echocardiogram.