Bayesian Survival Analysis: Cox Proportional Hazards-Model on Fatal Heart Attack

Mingchao Liu *
Department of Applied Math and Statistics
Johns Hopkins University
Baltimore, MD 21218
mliu122@jhu.edu

Jialin Guo
Department of Applied Math and Statistics
Johns Hopkins University
Baltimore, MD 21218
jguo81@jhu.edu

Abstract

Survival analysis is a method that looks at the duration of a time until one specific event happens. In short, "time for an event of interest". It is especially useful in clinical trails. Throughout the past few years, statisticians have worked on survival analysis and developed various models correspondingly.

In this paper, we look at Cox Regression Analysis, one of the most useful model in developing relationships between risk factors and target event. We apply Cox Regression analysis from Bayesian's approach on a heart-failure patient dataset. We analyze the effects of various risk factors on overall patient survival time; complete a table with a list of beta values, CI etc to check their significance.

Through sampling the posterior distribution with MCMC approach, we show the auto-correlation and trace plot of our resulted table and draws the baseline hazard function of the target risk factors and make further predictions about patients' survival time. Finally, we draw the conclusion that age, anaemia, CPK, high blood pressure, Serum Creatinine, those five factors play major role in patients' survival time and bring out some recommendations for future research.

1 Introduction to Survival Analysis

Survival analysis is one of the many popular topics in statistical world. What's important in survival analysis is "survival time", the time until an event of interest takes place. Such analysis is very useful in various fields like finance, economic, engineering and so on. Among all, it is most used in evaluating the effectiveness of certain treatments and predicting future diseases in medical and clinical fields. For example, We look at the time since a patient has been diagnosed with certain diseases until they died to predict future diseases; Or we record patients' characteristics at different time intervals to look for major factors that influence their survival time. In a nutshell, survival analysis is a powerful tool and major contributions in medical industry.

There are numbers of models in survival analysis. We have Log-rank tests that compare survival times between groups; Life tables, kaplan-meier curves that look at survival times of people in a group; And there is Cox regression, discover relationships between several risk factors and survival time. In this project, our major focus is Cox Proportional Hazard Model.

*Special thanks to instructor, Yanxun Xu, professor of Bayesian Statistics and all the TAs, Fall 2022, Johns Hopkins University

following the same style of 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

1.1 Literature Review for Bayesian Survival Analysis

Over the past few years, several statisticians have worked on survival analysis through a Bayesian approach. In *X.Zhao* et all, scholar Zhao and her colleagues look at the usefulness of the Weibull regression model in randomized censoring data. They assume each individual's life time data comes from Weibull distribution W(α,γ), and by letting μ=log(γ), they are able to obtains the likelihood function with distribution W(α,μ). Finally, they set a gamma prior and normal prior for α and γ to obtained the joint posterior distribution and sample it through Gibbs Sampler. They applied this model on a lymph sarcoma clinical data and analyze patient' relapse time. The results of their studies show that Weibull Distribution gives meaning statistics in clinical trails with randomized censoring data and proves the effectiveness of Gibbs Sampling in MCMC in computing complex posterior distribution.

In another paper written by scholar *D.Alveres* et all, his group introduce several of the most popular and useful Bayesian survival analysis models and presents a sample data analysis through R programming. They also brings detailed introduction to several important functions like *Survival Function* ($S(t)$) and *Hazard Function* ($H(t)$). They talk about accelerated failure time, proportional hazards model, mixture cure model, competing risks model and many others.

Finally, in *T.Soodejani* et all, his team makes a comparison of the Bayesian cox survival model and classical cox regression in determining significant risk factors. They apply their model fitting on a heart-failure patient dataset. Similarly, MCMC is also used when sampling posterior in Bayesian approach. Although two different approaches yield similar outputs, they conclude that the Bayesian model tends to be more accurate when effective sample size is limited.

In our project, we get some insights from the second and third paper. We deploy on the same dataset for heart-failure patients as Soodejani's group. We aim to conduct a more extensive data analysis and visualize the survival function and hazard function on Bayesian cox models with a prior of our own choice in an attempt to predict the risk of individuals dying from heart failure given certain circumstances.

2 Data and Research Question

2.1 Presentation of Data

This is a dataset originally collected from Faisalabad Institute of Cardiology. The time of collection is Apr till Dec in 2015. We obtains it from Kaggle. It contains about 299 patients and their personal information(target risk factor), follow up time(time until event of interest), their state(dead or alive). They are each categorized as having high cardiovascular risk (eg. they have high blood pressure, diabetes etc).

This is a brief overview of the 11 risk factors of patients (Figure1): Here is the distribution of the patient's death state. As we can see, majority of patients are still alive(death state = 0), only about 100 patients died during the experiment time(death state = 1). This could gives us potential bias in our fitted model since the censored data is large.

2.2 Important Definitions

- **Time to failure**: the length of time to event of interest. In our data, *follow-up time* represents the time to failure and *death state* represents event of interest.
- **Censoring**: when observation is halted or survival time of individual is not observed. One of the most common type is **right censoring**, it could be due to patient drops out or patient simply lives through the entire time span. In our data, *death state = 1* means it is censored.
- **Survival Function**($S(t)$): a function shows probability of a patient's survival after certain time.
- **Hazard Function**($H(t)$): the instantaneous rate of failure at a certain time when individual is still alive.
- **Risk factor**: certain triggering factors that may potentially increase prob of event of interest. In our dataset, there are **eleven** risk factors, that includes age, anaemia, CPK, diabetes etc.

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking
0	75.0	0	582	0	20	1	265000.00	1.9	130	1	0
1	55.0	0	7861	0	38	0	263358.03	1.1	136	1	0
2	65.0	0	146	0	20	0	162000.00	1.3	129	1	1
3	50.0	1	111	0	20	0	210000.00	1.9	137	1	0
4	65.0	1	160	1	20	0	327000.00	2.7	116	0	0

Figure 1: Overview of Data

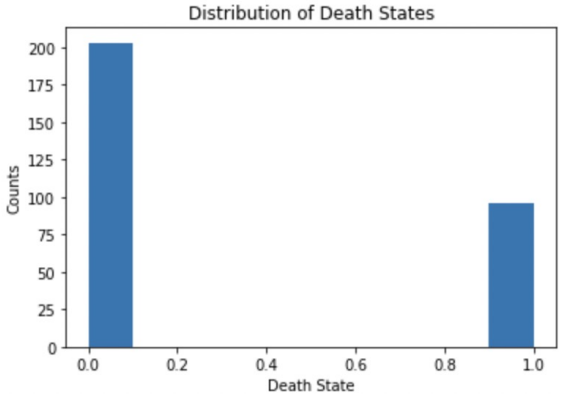

Figure 2: Distribution of Patient's Death State

2.3 Project Direction and Research Question

For our project, we will be focusing on **Cox Proportional Hazards Regression Analysis** in Bayesian's approach.

For our research question, we aim to look for **the effect of the 11 target risk factors on the patient's death due to heart failure**. To be more specific, which couple factors are more significant and which one is less important.

3 Method

3.1 Function Specification

In this project, since we want to look for the relationships between target risk factors and death event, we apply the **Cox Proportional Hazards-Model** under Bayesian's approach.

We pick ten equally length time interval that each contains 30days, and label them as λ_1, λ_2....λ_{10}. This gives the *baseline hazard function*($\lambda_0(t)$). As being said, it is a piece-wise constant function, where:

$$\lambda_0(t) = \lambda_i(i \in 1, 2, ...10)$$

With the *baseline hazard function*, we will then be able to derive the *hazard function*($h(t)$), which is:

$$h(t) = \lambda_0(t) * exp(X_i^T \beta)$$

Where X_i^T represents the design matrix of a combination of the eleven target risk factors($x_1, x_2, ...x_{11}$); And β consists eleven regression coefficients($\beta_1, \beta_2, ...\beta_{11}$) correspond to each risk factor x_i. [Note: An unique feature of cox regression model is that there is no beta interceptβ_0 here. It's included in λ_0.]

Another important function is *Survival Function*($S(t)$). It measures the probability of a patient's survival after certain time, hence can be expressed as :

$$S(t) = P(T > t)$$

where T represents to event of interest. There is a relationship between *Survival Function* and *Hazard Function*, it is:

$$h(t) = \frac{1}{S(t)} \lim_{\delta t \to 0} \frac{S(t+\delta t) - S(t)}{\delta t} = -\frac{S'(t)}{S(t)}$$

Building upon this, we can now derive a more illustrative *Survival Function* as:

$$S(t) = -exp\left(\int_0^t h(s)ds\right)$$

Besides above, we also compute the *Cumulative Hazard Function*($\Lambda(t)$), which is just an integration of *Hazard Function*($h(t)$):

$$\Lambda(t) = \int_0^t h(s)ds$$

And $S(t)$ can also be written as:

$$S(t) = -exp(\Lambda(t))$$

3.2 Parameters: Define Prior and Likelihood

Now, we define the priors for Baseline Hazard(λ_0) and regression coefficients(β), and compute the Hazard Function($h(t)$) and likelihood(L) of death correspondingly.

We denote the following priors:

- $\lambda_0 \sim \Gamma(10^{-2}, 10^{-2})$
- $\beta \sim N(\mu, \tau)$, where $\mu \sim N(0, 10^{-2})$ and $\tau = \sigma^{-2}$, as $\sigma \sim U(0, 10)$

Based on our priors, we can compute the Hazard Function($h(t)$):

$$h(t) = \lambda_0(t) * exp(X_i^T \beta)$$

Here, $h(t)$ is a 299x10 matrix that measures each individual patients hazard rate at ten different time intervals.

As for likelihood, it is very hard to find a closed form and straight forward function. So we first assumes a parameter μ, where:

$$\mu = \textit{Exposure} * h(t)$$

Here, *Exposure* is a 299x10 matrix that records the time that individual patient lives. For example, if Tom has lived for 50 days, that is ONE full time interval(λ_1) of 30 days and another 20 days in second time interval(λ_2). Then his *Exposure* matrix will be $[30, 20, 0, 0, 0, 0, 0, 0, 0, 0]$.

Through Matrix Multiplication, we obtain the parameter μ. We go on to assume likelihood(L) follows a *Poisson Distribution*. That is, we first create a new 299x10 matrix, *death* that contains only binary value $0, 1$. If a patient died on certain timer interval, say λ_3, then his or her death matrix will be $[0, 0, 1, 0, 0, 0, 0, 0, 0, 0]$.

Based of this, we can compute the likelihood: $L \sim Pois(\mu)$ and looks for the probability of observing each individual's death matrix in the distribution. That is for example, if Tom has a death matrix of $[0, 0, 1, 0, 0, 0, 0, 0, 0, 0]$, the prob of observing such matrix in a $Pois(\mu)$ shall be the likelihood.

3.3 Time Varying Effect

Beside Cox regression on Beta Coefficients, we also apply our Cox model on time varying effect. That is: we denote a function of Beta($B(t)$) that generates different Beta Coefficient(b_i) for each time interval(λ_i).

In this way, we derive a new *Hazard Function*($h(t)$):

$$h(t) = \lambda_0(t) * exp(X_i^T * B(t))$$

Under the new function, we are able to observe individual's survival function and hazard function at different time range.

3.4 MCMC Computation

We use the Metropolis-Hastings algorithm to implement the sampling analysis of the posterior distribution.

In the sampling process we use Poisson's zero trick to calculate likelihood. The prior distribution is mentioned above, and the normal distribution centered on the previous sample as the proposal function is implemented for sampling by using **pymc3**.

From the trace plot (Figure.4) , we can see that our MC-MC process converges successfully for each parameter. A lot of features is centering around 0, implying that they may not be that important. for the other parameters, their data distribution significantly deviates from 0, implying that the corresponding indicators of these parameters can have a significant impact on the recurrence of heart disease

By analyzing the auto-correlation plots (Figure.3) we get satisfactory low auto-correlation by using 20 as gap for the thinning process.

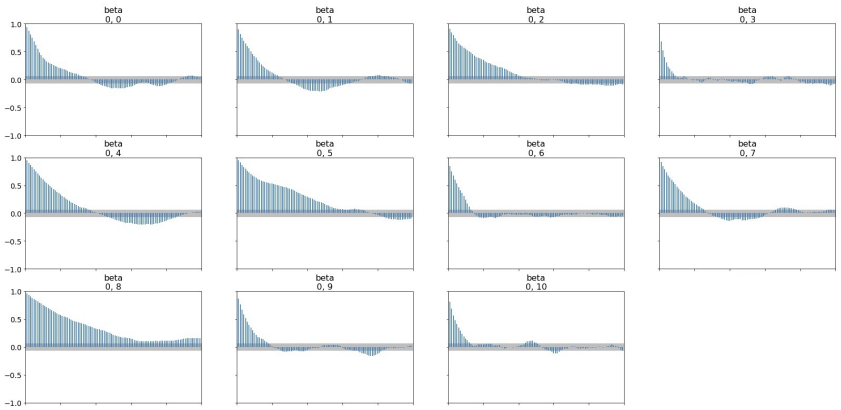

Figure 3: Autocorrelation of $\beta[0 \sim 10]$

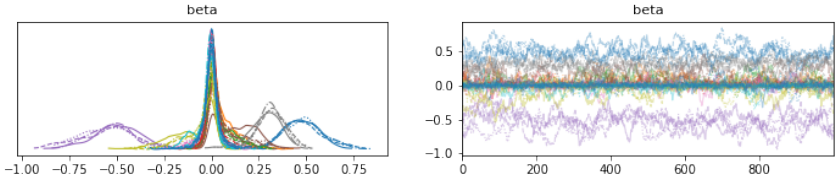

Figure 4: Traceplot of $\beta[0 \sim 10]$

4 Data Analysis and Discussion

4.1 Beta Coefficients

By putting in all eleven risk factors in X_i and running the four chains MCMC, we generate a chart of eleven beta coefficients(B_i) that contains their mean(μ_i), their Confidence Interval(CI) with $2.5\%, 50\%$ and $97,5\%$ correspondingly, and their probability of $P(B_i > 0)$.

To determine whether the Beta Coefficient is significant, we look for two criteria: 1.whether zero is includes in CI; 2. whether $P(B_i > 0) > 0.7$

Under such criteria, we conclude from the figure presented that: *age, anaemia, CPK, high blood pressure, Serum Creatinine*, those five factors play major role in affecting patients' survival time.

	mean	2.5%	50%	97.5%	Prob>0
age	0.481357	0.287372	0.477493	0.700739	1.00000
anaemia	0.047135	-0.037233	0.024862	0.219024	0.76375
creatinine_phosphokinase	0.038955	-0.052213	0.017859	0.218429	0.71075
diabetes	0.008507	-0.067888	0.004176	0.120325	0.56400
ejection_fraction	-0.535970	-0.805986	-0.527295	-0.284597	0.00000
high_blood_pressure	0.084125	-0.031688	0.045058	0.316033	0.82725
platelets	-0.010133	-0.155725	-0.004414	0.092417	0.43575
serum_creatinine	0.305092	0.119076	0.310242	0.447986	0.99875
serum_sodium	-0.061930	-0.309460	-0.025199	0.041008	0.24375
sex	-0.015280	-0.145306	-0.007722	0.064984	0.39425
smoking	0.000250	-0.098058	-0.000676	0.120212	0.48750

Figure 5: Table of $\beta[0 \sim 10]$

4.2 Cumulative Hazard Function and Survival Function

Here is some example of the output of the traditional Cox's model. (Figure. 6) Here we use the model to make predictions on Patient205 and patient183. Patient205's data is censored 187 days after her record begins, indicating she is recovering well. In the fig the green line is relatively lower in the Cumulative Hazard plot and higher in Survival functions, indicating a higher possibility of survival.

On the other hand, patient183 has a distinct higher hazard value and lower Survival rate. He actually died on the 162 day after being recorded.

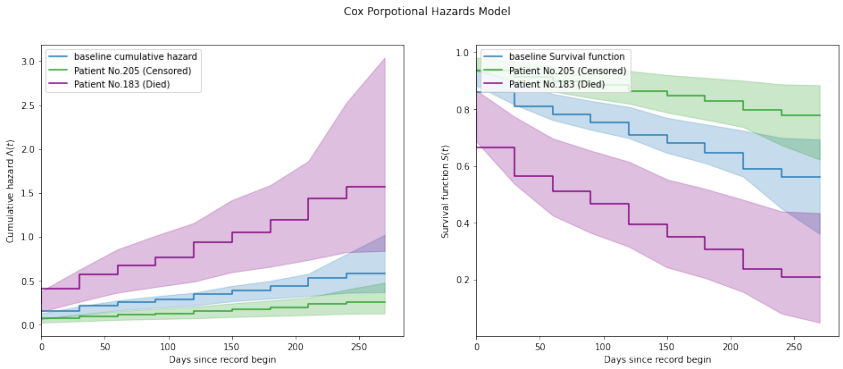

Figure 6: Cox Proportional Hazards Model

There's an obvious limitation: the output of our model is the same shape for all people, because the only thing that varies over time is the baseline hazard. The risk factors' influence stay constant.

Because the traditional Cox Model, time is only considered in the baseline hazard function, which is the blue line in Figure.6 all the predictions made by this model is baseline times a constant, in which the constant is decided by the risk factors. All the predictions are in the same shape, only differs in scale. It lacks of ability to utilize the fact that the influence of risk factors is actually varying over time. For example, a smoker's possibility of death is likely to grew over time as his Nicotine dependence gets stronger. In order for the model to also reflect this phenomenon, we want the coefficient for smoking to vary over time, getting bigger and bigger as time goes on. The fact is way more complicated than this assumption, but this is the intuition. To model this, instead of giving each risk factor one coefficient, the time varying cox model give 10 coefficients for each risk factor, indicating their effect on 10 arbitrary time stages. Figure 7 shows that how the coefficient for age changes over time. From the intuition that the coefficient is not likely to have a drastic change, we use a Gaussian random walk to model each series of the changing coefficients.

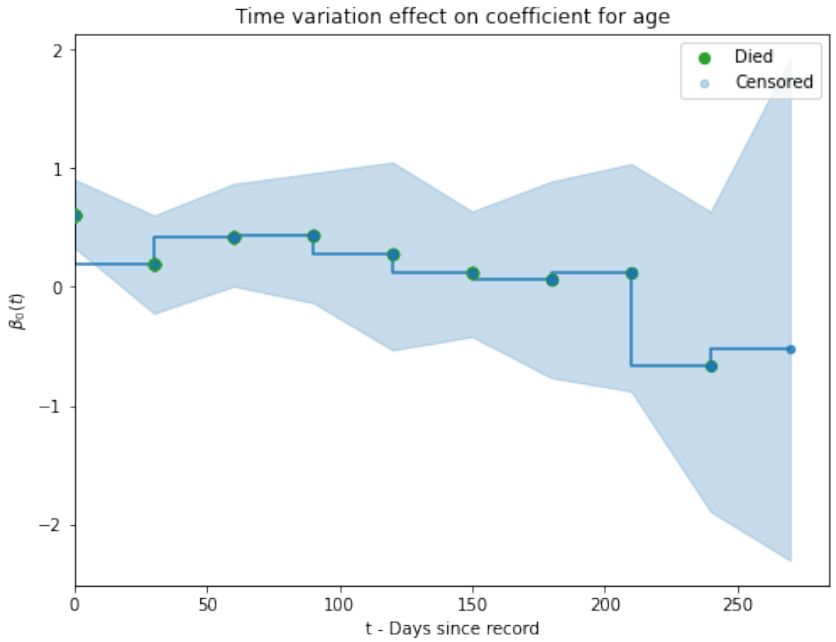

Figure 7: Time Variation of parameters

After making each coefficient "a series of coefficients" and implementing the Gaussian random walk to model the time varying effect, the new model can output Survival function of different shapes for different people given the fact that baseline hazard function is the same for everyone.

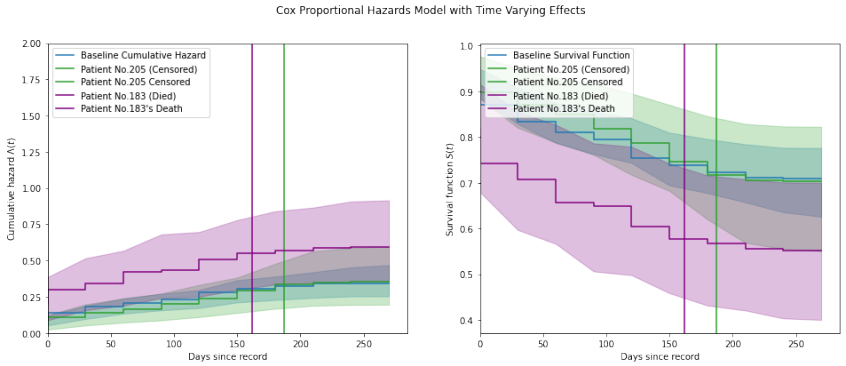

Figure 8: Cox Model with Time Varying effect

5 Conclusion

Overall, we have successfully demonstrated how to use Cox regression Analysis under Bayesian Approach to analyze relationships between risk factors and death event. Meanwhile, we also draw the cumulative hazard function, survival functions and those functions under time varying effect for patients of different states(dead or alive) to look for individual's survival rate. However, we do notice and acknowledge to certain problems and drawbacks inside our project.

First, we have limited amount of un-censored data. Only 1/3 of our patients died inside the ten times interval and major of them are censored. This could contribute potential bias in our model fitting process and potentially influence results. Second, originally as we conduct MCMC with Metropolis-Hastings, we hand-write each steps and loops. But our results does not converge and has extremely undesired auto-correlation plot. So we choose to turn to *Pymc3* method in Python. Last, as

we derive our likelihood function, we use *Poisson Distribution* and the *death* matrix. If a patient died in first period, all the other period will have 0 inside matrix. But we are unsure whether or not the zeros after a patient's death will still contribute to our model somehow. We only assume they won't in our computation.

Hence, for future reference and recommendations, we suggest to apply model on larger dataset that contain at least over 50% of uncensored data just to improve the accuracy of fitted model. Also, we will try to apply *Weibull Distribution* to replace *Poisson Distribution*.

References

[1] Alvares, D., Lázaro, E., Gómez-Rubio, V., & Armero, C. (2021). Bayesian survival analysis with BUGS. Statistics in Medicine, 40(12), 2975-3020.

[2] Hoff, P. D. (2009). A first course in Bayesian statistical methods (Vol. 580). New York: Springer.

[3] Zhao, X., Yu, C., & Tong, H. (2008, May). A Bayesian approach to weibull survival model for clinical randomized censoring trial based on MCMC simulation. In 2008 2nd International Conference on Bioinformatics and Biomedical Engineering (pp. 1181-1184). IEEE.

[4] Tolley, H. D., Barnes, J. M., & Freeman, M. D. (2016). Survival analysis. In Forensic Epidemiology (pp. 261-284). Academic Press.

[5] Martin, R. (2014). Stat 461–Applied Probability Models Some Lecture Notes.