Leukemia Cancer Treatment Survival Analysis from a Bayesian Perspective

William Qian

2024-05-12

Introduction

Leukemia cancer caused approximately 2.5% of all new cancer cases and 3.1% of cancer-related mortality (Huang et al. (2022)), tons of researches have been conducted to improve the prognosis of leukemia cancer patients, and various of treatment strategies have been developed to prolong the survival time of patients. One of the most important tasks in leukemia cancer research is to test whether a new treatment is effective in improving the survival time of patients. Survival analysis, which estimates the time until the occurrence of an event of interest, such as death, is crucial for testing the efficacy of new treatments and identifying prognostic factors that affect survival outcomes.

The Cox model (Breslow (1975)), a widely used statistical method for survival analysis, is based on the proportional hazards assumption and provides valuable insights into the relationship between covariates and survival outcomes. However, traditional frequentist approaches to survival analysis with the Cox model have limitations, such as restrictive assumptions and difficulties in handling complex data structures.

This analysis used method that seeks to extend the traditional Cox model with Bayesian perspectives, which incorporated hierarchical priors that concluded from real world researches, improving model flexibility and reliablity (Muehlemann et al. (2023)). The Bayesian Cox model was applied to a leukemia cancer treatment dataset to demonstrate its advantages in survival analysis, including more accurate parameter estimation, better model fit, and more informative inference.

In this analysis, we first analyze the structure of our data and then introduce the Bayesian framework and Cox model of its theoretical foundation. After that, a detailed mathmatical deduction was used to find form of the posterior distribution of the parameters. We then utilize the Markov Chain Monte Carlo (MCMC) algorithms with Stan to draw posterior samples and estimate the parameters of the Bayesian Cox model. A detailed interpretation of the model is also provided to make the inference.

In conclusion, this work implemented a Bayesian Cox model by extending the traditional Cox model with hierarchical priors, providing a more flexible and reliable approach to survival analysis. The results of the analysis provides a more accurate prediction of survival outcomes and a better understanding of the relationship between covariates and survival time in patients with different treatments.

Data Structure and Notation

The dataset used in this analysis is the leukemia cancer dataset from Kaggle (Djegou (2022)). This dataset is collected from a research that test the efficacy of a new treatment for leukemia cancer patients, and includes information on treatment type and survival outcomes.

Part of the data is shown below:

Table 1: Partial data of the leukemia cancer dataset

id	t	event	treatment
1	6	1	1
2	6	1	1
3	6	1	1
4	7	1	1
5	10	1	1
6	13	1	1
7	16	1	1
8	22	1	1
9	23	1	1
10	6	0	1

We can see that there are 42 observations in the dataset.

The `id` variable is the unique identifier for each patient, which is not relevant to the analysis and will be removed.

The `treatment` variable, is the main target we care about. There are two values in this variable, `0` and `1`, where `0` is the placebo group and `1` is the treatment group. Our goal is to find out whether the new treatment is effective in improving the survival time of patients compared to the standard treatment.

The `event` variable consists of two levels, `0` and `1`, where `0` indicates that the patient is censored and `1` indicates that the patient has died. This variable is particularly important in survival analysis, as it indicates whether the patient has experienced the event of interest. And in the following analysis, we will have different treatment for censored and uncensored (dead) data.

The `t` variable is the survival time of the patients, which is the time from diagnosis to death or censoring. The unit of the time is in weeks.

Characteristic	**0**, N = 21	**1**, N = 21	**p-value**
t	8 (4, 12)	16 (9, 23)	0.004
event	21 (100%)	9 (43%)	<0.001

In order to perform survival analysis, we need to define some notation.

- Let t_i be the survival time of patient i or the time at which patient i is censored.
- Let d_i be the event indicator of patient i, where $d_i = 1$ if patient i has died and $d_i = 0$ if patient i is censored.
- Let x_i be the covariates of patient i, in our analysis, it is specified for the `treatment` variable, where $x_i = 1$ if patient i is in the treatment group and $x_i = 0$ if patient i is in the control group.
- And we use the term $\mathcal{D}$ to represent the dataset, when we talk about the censored data, we use $\mathcal{D}_{d_i=0}$ to represent the censored data, and $\mathcal{D}_{d_i=1}$ to represent the uncensored data.

Methodology

Bayesian Framework

The Bayesian framework provides a flexible and intuitive approach to statistical modeling by incorporating prior information and updating it with observed data to obtain the posterior distribution of the parameters.

To discuss probabilities related to θ given the data y, we must start with a model that offers a joint probability distribution for both θ and y. This joint distribution is typically expressed as the product of two components: the prior distribution $p(\theta)$, and the likelihood or sampling distribution $p(y|\theta)$. The mathematical representation is:

$$p(\theta, y) = p(\theta)p(y|\theta)$$

By applying the principle of conditional probability, specifically Bayes' rule, and conditioning on the observed data y, we derive the posterior density:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}$$

Here, $p(y)$ is the marginal probability of y, computed as $\sum_\theta p(\theta)p(y|\theta)$ for discrete θ, or as an integral $\int p(\theta)p(y|\theta)d\theta$ for continuous θ. We can also express the posterior density in

a proportional form by excluding the constant term $p(y)$, as it does not depend on θ and remains constant for a given y:

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

Note that the term $p(y|\theta)$ should be taken as a function of θ instead of y, in practice, we usually use the likelihood function, which is the probability of observing the data y given the parameter θ.

The Cox Model

The proportional hazards model, also known as the Cox model, is a widely used statistical method for survival analysis. The Cox model is a semi-parametric model that estimates the hazard function, which is the instantaneous rate of failure at time t given that the individual has survived up to time t. The Cox model assumes that the hazard function is a product of a baseline hazard function and an exponential function of covariates. The mathematical representation of the Cox model is:

$$h(t|x) = h_0(t) \exp(\beta^T X)$$

Traditionaly, to estimate the parameters $\beta = (\beta_1, \beta_2, \dots, \beta_p)$, we can use maximum likelihood estimation (MLE) to maximize the likelihood function and obtain those parameters. However, the MLE has some limitations, such as restrictive assumptions.

Generally, we care more about the ratio of the hazard functions between two groups, which is the hazard ratio, that is why $h_0(t)$ is not that important. However, sometimes we do want to know the baseline hazard function, which is the hazard function when all the covariates are 0. This is wehn the Nelson-Aalen estimator comes in handy. The Nelson-Aalen estimator is a non-parametric estimator of the cumulative hazard function. The Nelson-Aalen estimator is defined as:

$$H_0(t) = \sum_{j=1}^{k} \frac{d_j}{\sum_{l \in R_j} exp(\beta^T (x_l - x^*))}$$

Where d_j is the number of events at time t_j, R_j is the risk set at time t_j, and x^* is the covariates value in reference group. The survival function can then be expressed as:

$$S(t) = \exp(-H_0(t) exp(\beta^T x))$$

Note that $S(t)$ is exactly what we care about in the survival analysis, which is the probability that the event has not happened at time t.

Adjusted Bayesian Cox Model

Combing the two concept shown above, we can now have the Bayesian Cox model. The idea is very intuitive, we can use the Bayesian method to estimate the parameters of the Cox model.

However, we are facing some challenges when applying the Bayesian method to the Cox model. The likelihood function of the Cox model is intractable due to the presence of the baseline hazard function, which is not specified in the model. And we have two sets of data that one group is censored and the other is not. Those two different group of data should be treated differently in the model, which is not properly handled in the original Cox model.

To address these challenges, we propose an adjusted Bayesian Cox model.

Estimate the Baseline Hazard Function

First of all, let's talk about the $H_0(t)$, the baseline hazard function. The traditional method is to use Nelson-Aalen estimator, however, this estimator is too complicated to implement in the Bayesian framework. Thinking about the essence of the baseline hazard function, we can find that the baseline hazard function should have a attribute that is independent of the covariates, and also should be a monotonically increasing function, just like the Nelson-Aalen estimator. So we can use a simple exponential function to represent the baseline hazard function, which is:

$$h_0(t) = \exp(\eta)$$

The biggest advantage of this representation is that we can now estimate the baseline hazard function in the Bayesian framework, and to be noticed that the form of $exp(\eta)$ can be easily integrated into the rest part of the model $exp(\hat{\beta}^T x)$, making η actually the intercept of the linear model.

Uncensored Data

The uncensored data, which is the data that the event has happened. Since our baseline hazard function is now a constant, the hazard function of the uncensored data can be expressed as:

$$h(t|x, d_i = 1) = \exp(\eta + \beta^T x)$$

To implement the Bayesian method, we need to find the PDF, which is $f(t|x, d_i = 1)$, that is:

$$\begin{aligned} f(t|x, d_i = 1) &= h(t|x, d_i = 1) S(t|x, d_i = 1) \\ &= \exp(\eta + \beta^T x) \exp(-\exp(\eta + \beta^T x) t) \end{aligned}$$

Then, we could find the likelihood function of the uncensored data, which is:

$$\begin{aligned} L(\mathcal{D}_{d_i=1}|\eta,\beta) &= \prod_{i=1} f(t_i|x_i, d_i = 1) \\ &= \prod_{i=1} \exp(\eta + \beta^T x_i) \exp(-\exp(\eta + \beta^T x_i) t_i) \end{aligned}$$

Censored Data

For the censored data, things are a little bit different. The censored data is the data that the event has not happened, which means our patients are alive. The event time of the censored data is not the actual event time, but the time that the patient is still alive, which meaning that the event time t is nor known, so we could not obtain the PDF $f(t|x, d_i = 0)$ like what we did in the uncensored data. Instead, we work with the survival function $S(t|x, d_i = 0)$, or its complementary cumulative distribution function (CCDF) to handle these cases.

In our analysis, we only have right censored data. Meaning that for the given time t, we only know that the event is not happened at t. Therefore, the likelihood contribution of this data point is based on the probability that the event time T is greater than t, meaning $S(t|x) = P(T > t|x)$. Then, the likelihood function of the censored data is:

$$\begin{aligned} L(\mathcal{D}_{d_i=0}|\eta,\beta) &= \prod_{i=1} S(t|x, d_i = 0) \\ &= \prod_{i=1} \exp(-\exp(\eta + \beta^T x_i) t_i) \end{aligned}$$

MCMC Sampling

To estimate the posterior distribution of the parameters in the Bayesian Cox model, we use Markov Chain Monte Carlo (MCMC) algorithms to draw samples from the posterior distribution. The MCMC algorithms generate a sequence of samples that converge to the target distribution, allowing us to estimate the posterior distribution of the parameters. In this analysis, we use the Stan software to implement the MCMC sampling and estimate the parameters of the Bayesian Cox model.

Computation

Posterior Derivation

Consider what parameters we need to estimate in the model, we have η and $\beta_{treatment}$. We can firstly set prior for those parameters. Let's say both of the parameters follows a normal

distribution that:

$$\eta \sim N(0, 10)$$
$$\beta_{treatment} \sim N(0, 10)$$

Then, considering of the likelihood function of the uncensored data and censored data, we can have the posterior distribution of the parameters as:

$$\begin{aligned} p(\eta, \beta_{treatment}|\mathcal{D}) &\propto p(\mathcal{D}|\eta, \beta_{treatment})p(\eta)p(\beta_{treatment}) \\ &\propto L(\mathcal{D}_{d_i=1}|\eta, \beta)L(\mathcal{D}_{d_i=0}|\eta, \beta)p(\eta)p(\beta_{treatment}) \\ &\propto \prod_{i=1;d_i=1} \exp(\eta + \beta_{treatment}x_i) \exp(-\exp(\eta + \beta_{treatment}x_i)t_i) \\ &\cdot \prod_{i=1;d_i=0} \exp(-\exp(\eta + \beta_{treatment}x_i)t_i) \\ &\cdot exp(-\eta^2)exp(-\beta^2_{treatment}) \end{aligned}$$

Metropolis-Hastings Algorithm

For this particular case, I would suggest using Metroplis-Hastings algorithm to sample the posterior distribution of the parameters.

To implement the MH algorithm, we need prepare the following equations: $h(t|x), S(t|x), L(\mathcal{D}_{d_i=0}|\eta, \beta)$, and $L(\mathcal{D}_{d_i=1}|\eta, \beta)$ and the priors of the parameters, which are $\beta_{treatment} \sim N(0, 10)$ and $\eta \sim N(0, 10)$. Then we can start the sampling process.

The first step is the initialization of the parameters, we can set $\eta = 0$ and $\beta_{treatment} = 0$. We should also choose proposal distribution variances for the parameters, which could be $\sigma_\eta = 1$ and $\sigma_{\beta_{treatment}} = 1$.

The following step is the main interation of the MH algorithm. Let's say we will iterate the algorithm for 5000 times. In each iteration, we will sample the new parameters from the proposal distribution. For example, in the iteration number k, for η, we will first sample a new value η^* from the normal distribution $N(\eta^{(k-1)}, \sigma^2_\eta)$. Then, we will compute the acceptance ratio:

$$r = \frac{p(\eta^k|\mathcal{D})}{p(\eta^{(k-1)}|\mathcal{D})} \cdot \frac{q(\eta^{(k-1)}|\eta^k)}{q(\eta^k|\eta^{(k-1)})}$$

Where q is the proposal distribuiton here. And since we have the acceptance ratio, we can then decide whether to accept the new value or not by comparing it to a random number from the uniform distribution. If we accept the new value, then $\eta^{(k)} = \eta^*$, otherwise, $\eta^{(k)} = \eta^{(k-1)}$.

We will do the same thing for $\beta_{treatment}$ in the same iteration.

And after the 5000 iterations, we will have 5000 samples for each parameter, usually, we will set a 1000 iterations as the burn-in period, which means we will discard the first 1000 samples, and then we will have 4000 samples left. In practice, we shall run several chains to make sure the convergence of the algorithm, and if we confirm the convergence, we can then combine the samples from different chains to get the final result, usually we can use the mean value of the chain as our estiamtes.

Data Analysis

Estimate the Parameters with Stan

Now, let's implement the adjusted Bayesian Cox model using the Stan to estimate the parameters of the model.

In our case, we will set the iteration number to be 5000, and the number of chains to be 5. After the sampling process, we obtained the following results.

As we can see in the table, the mean of the posterior distribution of the $\beta_{treatment}$ is -1.556461, which has a standard deviation of 0.4093681, and the credible interval is (-2.373586, -0.7818435). The mean of the posterior distribution of the η is -2.184783, which has a standard deviation of 0.2187248, and the credible interval is (-2.635452, -1.7754279).

Table 3: Summary of the posterior distribution of the parameters

	mean	se_mean	sd	2.5%	97.5%
beta_treatment	-1.560972	0.0056122	0.4074366	-2.387561	-0.7962161
eta	-2.181942	0.0031971	0.2213597	-2.631265	-1.7686307
lp___	-109.569255	0.0133302	0.9989581	-112.244320	-108.5861071

Now, let's check the convergence of the MCMC algorithm by plotting the traceplot of the parameters and the posterior distribution of the parameters. As we can see in the plot that all of the 5 chains are well mixed and converged, which means the process converged well.

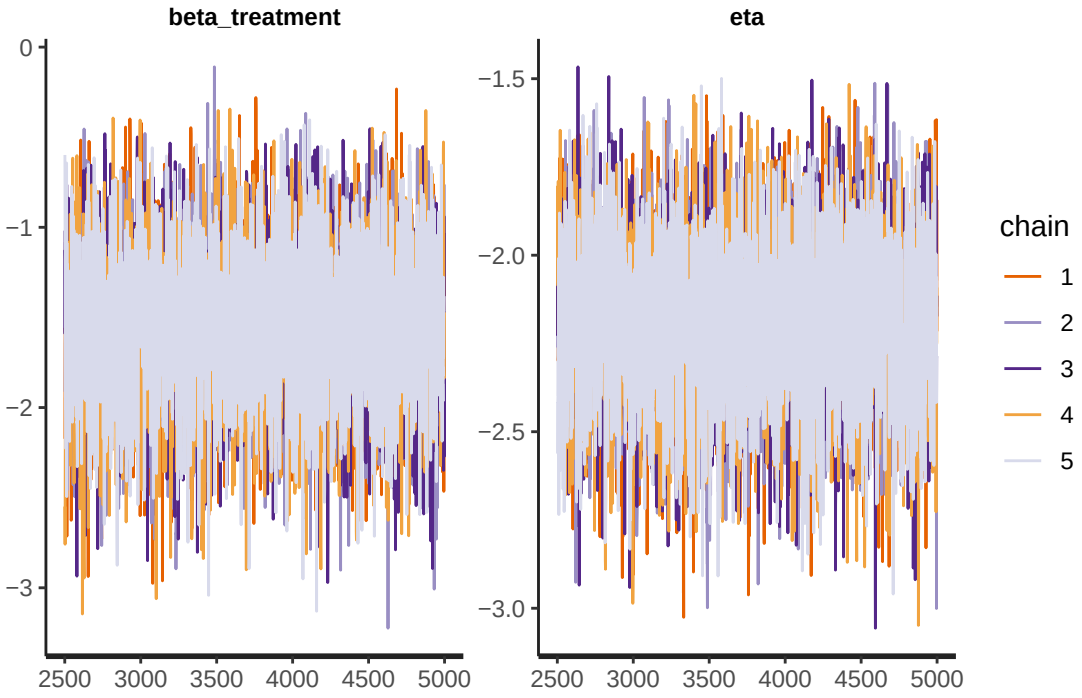

Figure 1: Traceplot of the parameters

We could also plot the distribution of the draws to check the result. We may find that the draws seems to be a normal distribution and condencd on our mean value, which is a good sign that the MCMC algorithm works well.

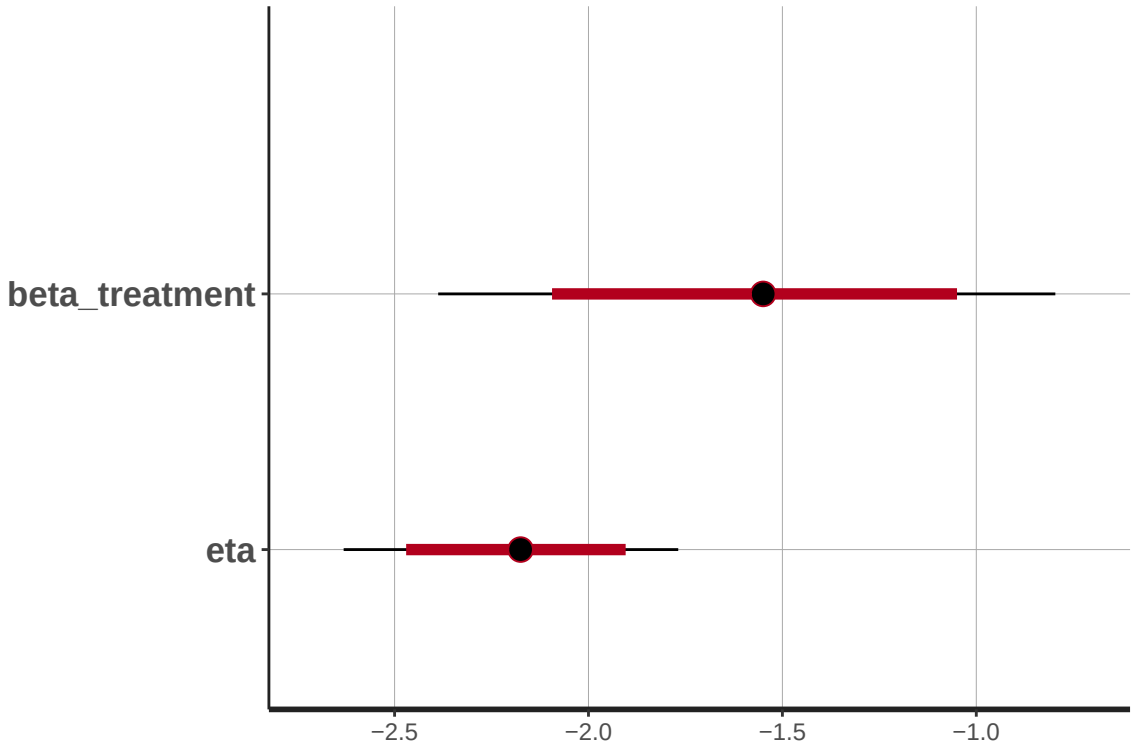

Figure 2: Posterior distribution of the parameters

To be noticed that, both of the distribution of η and $\beta_{treatment}$ are normal distribution, and neither of their credible intervals contains 0, which suggests those coefficients are significant in the model.

Now, with the posterior distribution of the parameters, we can plot the surival function of the treatment group and the placebo group. Recall that the survival function is defined as

$$S(t) = \exp(-\exp(\eta + \beta^T x)t)$$

So for the treatment group, the survival function is:

$$\begin{aligned} S(t) &= \exp(-\exp(\eta + \beta_{treatment})t) \\ &= \exp(-\exp(-2.184783 - 1.556461)t) \end{aligned}$$

And for the placebo group, the survival function is:

$$\begin{aligned} S(t) &= \exp(-\exp(\eta)t) \\ &= \exp(-\exp(-2.184783)t) \end{aligned}$$

The following plot shows the survival function of the treatment group and the placebo group. As we can see, the survival probability of the treatment group is higher than the placebo group, and the survival rate of the placebo group decreases faster than the treatment group, and decrease to 0 at around 40 weeks, while the treatment group still has a survival rate over 0.037 at that time, and the treatment group has a survival rate over 0.25 until the end of the research. This result indicates that the new treatment is effective in improving the survival time of patients.

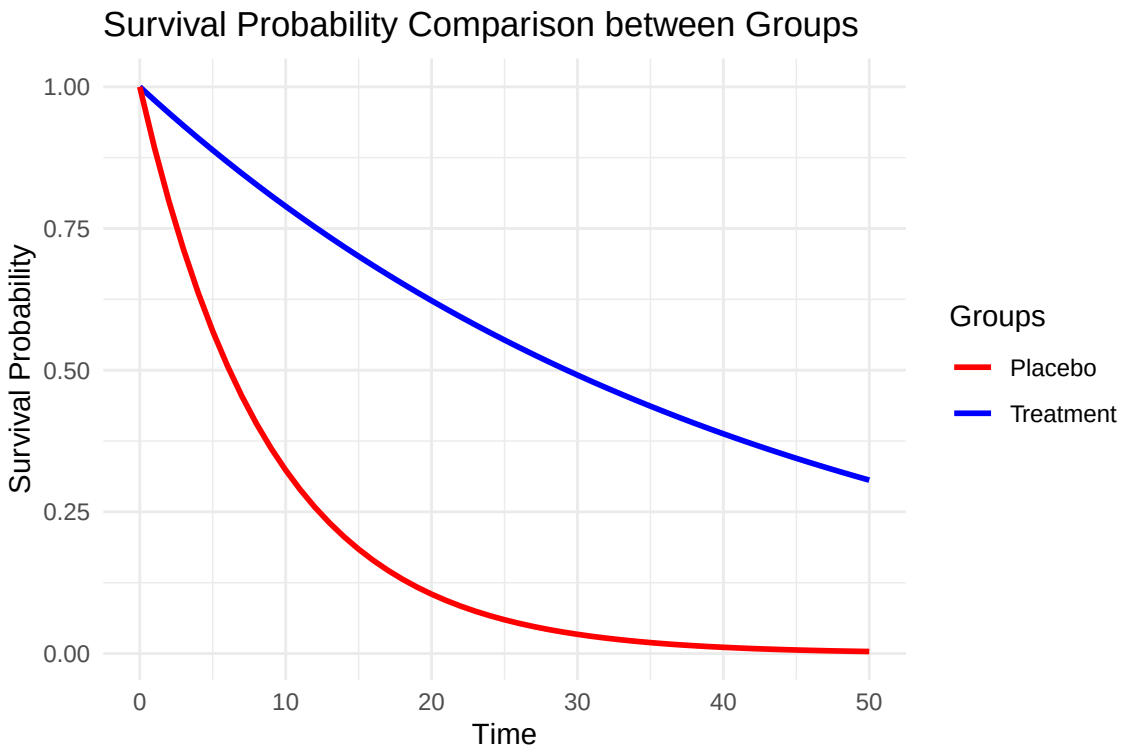

Figure 3: Survival Probability Comparison between Groups

Model Comparison

In order to check whether our model is effective in predicting the survival time of patients, we can compare the Bayesian Cox model with the traditional Cox model. We will use `survival` package in R as a reference to check the result.

We may see form the table that the result of the traditional Cox model is very similar to our adjusted Bayesian Cox model, which is a good sign that our model works well.

Table 4: Summary of the traditional Cox model

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
treatment	-1.572125	0.2076035	0.4123967	-3.812167	0.0001378

A survival plot is also provided to compare the survival probability of the treatment group and the placebo group fitted by the `survival` package. As we can see, the trend of survival probability also went very similar to our model.

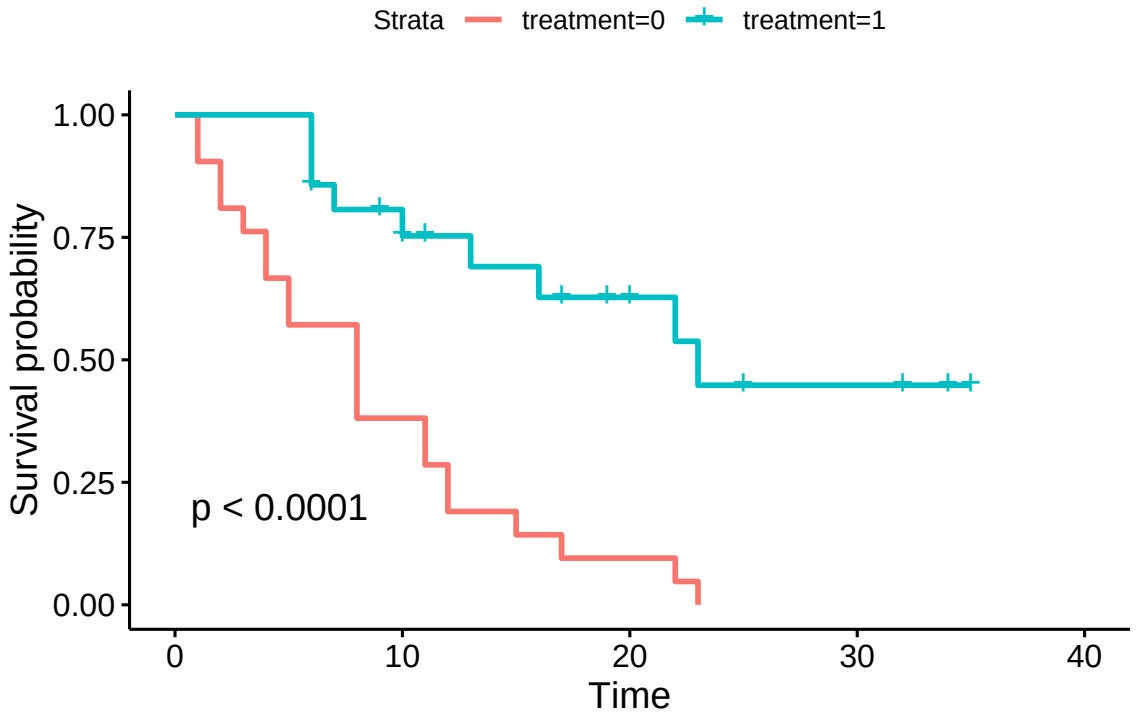

Figure 4: Survival Probability fitted by the Survival Package

Result Interpretation

Now we have been confirmed that our model is working well and the new treatment is effective in improving the survival time of patients. We can dive deeper into the result and make some interpretation.

Let's start with the interpretation of η. η is the parameter of the baseline hazard function, which could be also understand as the intercept of the linear model in the later part of the model. The mean of the posterior distribution of η is -2.184783, which means that the baseline

hazard function is $exp(-2.184783) = 0.112$. Note that the hazard function is constant across time, which means that the hazard function is 0.112 for all the patients.

Note that the Cox model, is also called the proportional hazard model, so a good way to interpret the $\beta_{treatment}$ is to interpret it as the hazard ratio. The mean of the posterior distribution of $\beta_{treatment}$ is -1.556461, which means that the hazard ratio of the treatment group to the placebo group is $exp(-1.556461) = 0.210$. This means that the hazard of the treatment group is 0.210 times the hazard of the placebo group, which means that the survival time of the treatment group is $1/0.210 = 4.76$ times longer than the placebo group.

References

The Github repository of this analysis can be found here: https://github.com/dukechain2333/PHP2530-Fianl-Paper

Breslow, N. E. 1975. "Analysis of Survival Data Under the Proportional Hazards Model." *International Statistical Review / Revue Internationale de Statistique* 43 (1): 45–57. https://doi.org/10.2307/1402659.

Djegou, Emmanuel. 2022. "Leukemia Cancer Data for Survival Analysis." https://www.kaggle.com/datasets/em survival-data-format.

Huang, Junjie, Sze Chai Chan, Chun Ho Ngai, Veeleah Lok, Lin Zhang, Don Eliseo Lucero-Prisno, Wanghong Xu, et al. 2022. "Disease Burden, Risk Factors, and Trends of Leukaemia: A Global Analysis." *Frontiers in Oncology* 12 (July): 904292. https://doi.org/10.3389/fonc.2022.904292.

Muehlemann, Natalia, Tianjian Zhou, Rajat Mukherjee, Munshi Imran Hossain, Satrajit Roychoudhury, and Estelle Russek-Cohen. 2023. "A Tutorial on Modern Bayesian Methods in Clinical Trials." *Therapeutic Innovation & Regulatory Science* 57 (3): 402–16. https://doi.org/10.1007/s43441-023-00515-3.