Bayesian Survival Analysis for Lung Cancer Data

Yuan Gao

May 12, 2025

1 Introduction

Survival analysis is a statistical method that analyzes time-to-event data, where the "event" could be failure, recovery, or any defined outcome of interest. Compared to traditional regression models, survival analysis takes censored observations into account, meaning that the event may not be observed within the study period. The survival and hazard functions are two important components; the former estimates the probability of surviving beyond a given time, and the latter describes the instantaneous risk of the event occurring. Common models include the Kaplan–Meier estimator, the Cox proportional hazards model, and parametric models such as Weibull or log-normal. Today, survival analysis is widely used in different fields like medical research, engineering, and social sciences.

Bayesian statistic offers us the opportunity to assess a parameter given a specific data set. It regards parameters as random variables and sets the observed data fixed. It can incorporate prior knowledge with observed data to update beliefs about unknown parameters using Bayes' theorem. The outcomes produced by Bayesian inference are called posterior distributions, which can quantify uncertainty. It is a useful method that is required in many specific fields.

The application of Bayesian approaches to survival analysis makes the process more flexible and informative, which can be an important tool in clinical trials.

In Section 3, I will introduce the main functions used in survival analysis and derive their relationship and the concept of censored data will be explained in Section 4. Then, Section 5 will introduce the Kaplan-Meier Survival Curve, which is a non-parametric estimators for survival functions. Besides, the log-rank test and graphical illustrations will be provided for understanding. Sections 6 and 7 will delve into semi-parametric and parametric approaches, respectively. In Section 6, I will focus on the Cox Proportional Hazards Model, where the derivation of maximal likelihood estimation will be included. Section 7 will extend the discussion to parametric survival models such as exponential, Weibull, and log-logistic distributions, along with the model comparison criteria: the Akaike Information Criterion (AIC). In Section 8, I will move on to the Bayesian statistics, and topics such as Bayes' Theorem, conjugate priors, and Markov Chain Monte Carlo (MCMC) methods will be formally introduced. Finally, Section 9 will synthesize the theoretical and methodological content by applying Bayesian survival analysis to practical examples using R. I will specify a model, make posterior inference, and interpret the results based on the R output.

1.1 Introduction to the data used in this project

In the practical section of the project, I will use the `lung` data set available from the survival package in R. The data contains subjects with advanced `lung` cancer from the North Central Cancer Treatment Group [1]. It includes the following 10 variables:

- **inst**: Institution code
- **time**: Survival time of a patient (days)
- **status**: Censoring status (1 = censored, 2 = dead)
- **age**: Age of a patient (years)

- **sex**: Male = 1, Female = 2
- **ph.ecog**: ECOG performance score as rated by the physician (0 = asymptomatic, 1 = symptomatic but completely ambulatory, 2 = in bed <50% of the day, 3 = in bed >50% of the day but not bedbound, 4 = bedbound)
- **ph.karno**: Karnofsky performance score (bad = 0, good = 100) rated by physician
- **pat.karno**: Karnofsky performance score (bad = 0, good = 100) rated by patient
- **meal.cal**: Calories consumed at meals (days)
- **wt.loss**: Weight loss in the last six months

The Karnofsky Performance Score is a standardized scale used to measure a cancer patient's overall functional status and ability to carry out daily activities [2].

2 Introduction to Survival Analysis

The development of survival analysis can be traced back to the 17-18 centuries. When dealing with demographic problems, the concept was introduced to study population longevity and mortality rates. In 1662, John Graunt, who was considered the father of demography, published *Natural and Political Observations* and proposed the method of calculating the mortality rate. Then in 1693, the famous British astronomer Edmond Halley firstly constructed a complete life table to estimate survival probabilities for different age groups. In the 18-19 centuries, Benjamin Gompertz proposed the Gompertz Life Table Model in 1825, which found out that there is an exponential correlation between mortality rate and age. Furthermore, in 1860, William Makeham formed the Gompertz-Makeham formula by adding an age-independent mortality component to the original Gompertz model. During the 19th century, with the introduction of t-distribution and probability theory, the method could be used for more complex statistical models. After stepping into the 20th century, Edward L. Kaplan and Paul Meier proposed the Kaplan-Meier (KM) estimator to estimate the survivor function in 1958, which effectively handled censored data. Besides, in 1972 David Cox developed the Cox Proportional Hazards Model to check whether other potential covariates would influence the survival time. Nowadays, with the help of programming, complex calculations can be dealt with well. [3]

3 Survivor Function and Hazard Function

3.1 Definition

In order to express the probability that a person survives longer than a certain time t, the survivor function: $S(t)$ is often used [4], mathematically, it can be expressed as:

$$S(t) = P(T > t) = 1 - F(t)$$

where:
- T is the random variable representing the survival time,
- $F(t) = P(T \leq t)$ is the cumulative distribution function (CDF) of T, which gives the probability that the survival time is at most t,
- $S(t)$ gives the probability that the survival time exceeds t.

There is also another function called the hazard function: $h(t)$. The hazard function represents the instantaneous potential per unit of time for the event to occur, conditioned on the event not occurring until t [4].

It can be expressed as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

3.2 Derivation of Survivor Function and Hazard Function

Since we know that the hazard function is:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

and by using conditional probability, we have that:

$$P(t \leq T < t + \Delta t \mid T \geq t) = \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)}$$

with $S(t) = P(T > t)$, the above equation then becomes:

$$P(t \leq T < t + \Delta t \mid T \geq t) = \frac{P(t \leq T < t + \Delta t)}{S(t)}$$

Then we have $h(t)$:

$$h(t) = \lim_{\Delta t \to 0} \frac{\frac{F(t+\Delta t) - F(t)}{S(t)}}{\Delta t}$$

It also be represented by:

$$\lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} = f(t),$$

where $f(t)$ denotes the derivative of $F(t)$ if the density is well defined. Then $h(t)$ equals to:

$$h(t) = \frac{f(t)}{S(t)}$$

To do further calculations, the relationship between $f(t)$ and $S(t)$ is needed:

$$F(t) = P(T \leq t)$$

So,

$$S(t) = P(T > t) = 1 - F(t)$$

Also:

$$f(t) = \frac{dF(t)}{dt}$$

We have the following:

$$\frac{dS(t)}{dt} = \frac{d}{dt}(1 - F(t)) = -\frac{dF(t)}{dt} = -f(t)$$

Since

$$h(t) = \frac{f(t)}{S(t)}$$

we have

$$f(t) = h(t) \cdot S(t)$$

So,

$$-\frac{dS(t)}{dt} = h(t)S(t)$$

$$\frac{dS(t)}{S(t)} = -h(t)dt$$

Integrate both sides, we have:

$$\int_0^t \frac{dS(t)}{S(t)} = -\int_0^t h(t)dt$$

Thus,

$$\ln S(t) = -\int_0^t h(t)dt$$

We furthermore take the exponential of both sides:

$$S(t) = e^{-\int_0^t h(u)\,du}$$

So, we eventually get the formula for the survivor function, which is:

$$S(t) = e^{-\int_0^t h(u)\,du}$$

For $t = 0$ we have $S(0) = e^{-0} = 1$ which means the probability of the participants to survive at time 0 is 1.
Specifically, the cumulative hazard function $H(t)$ is defined as the integral of the hazard function $h(t)$ over time:

$$H(t) = \int_0^t h(u)du$$

From:

$$S(t) = e^{-\int_0^t h(u)\,du}$$

we get:

$$S(t) = e^{-H(t)}$$

So, the survivor function is completely determined by the cumulative hazard function.

3.3 Examples of Survivor Function under different Hazard Function

We know that the correlation between the survivor function and the hazard function is:

$$S(t) = e^{-\int_0^t h(u)\,du}$$

For the Weibull distribution, the corresponding hazard function is

$$h(t) = \lambda k t^{k-1}$$

Where λ is a scale parameter (related to the rate of failure) and k is a shape parameter. With different parameters λ and k, the function behaves differently.

We now explore how different values of the parameters will affect the shape of the hazard and survivor functions.

If $\lambda > 0$ and $k > 1$, it is a increasing function.
If $k > 1$, the hazard function increases with time, which means that the item becomes more likely to fail as it ages. From $h(t) = \lambda k t^{k-1}$ we can get the survivor function:

$$S(t) = e^{-\lambda t^k}$$

From this formula, we know that the probability of survival decreases more rapidly over time as the hazard function increases, and for larger k, the probability of survival decreases more dramatically. The graphs below show the trend of the increasing hazard function and related survivor function.

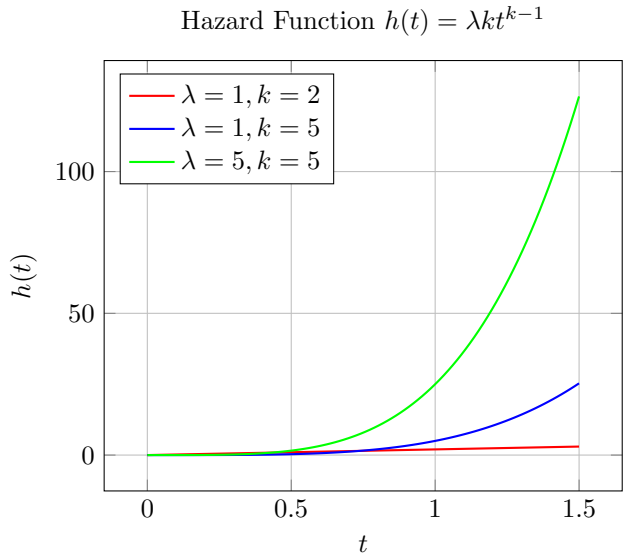


Figure 1: Hazard Function for Different λ and k

From this graph we find that:

- Higher k leads to a more dramatic increase in the failure rate.
- When k increases, the hazard function grows non-linearly, meaning the failure rate accelerates as time progresses.
- Higher λ amplifies the overall failure rate and shifts the hazard function upward, meaning failure occurs at a much higher probability at all times.
- Systems with large k and λ fail extremely fast. The green curve with $k = 5$, $\lambda = 5$ shows that early survival is possible, but as time progresses, the system rapidly deteriorates.

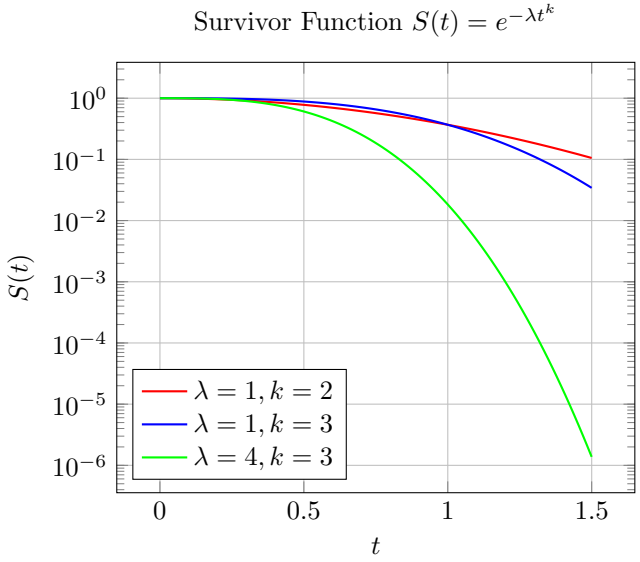


Figure 2: Survivor Function for Different λ and k

From this graph we find that: higher k leads to sharper survival drops, and higher λ results in a steeper decline. When looking at the green curve, we find that the combination of high k and high λ leads to near-instant failure.

If $\lambda > 0$ and $0 < k < 1$, it is a decreasing function.
The hazard function decreases with time, which means that the highest failure rate occurs at the beginning, then the system is more likely to fail early, and those that survive become more stable over time. From $h(t) = \lambda k t^{k-1}$ we can get the survivor function:

$$S(t) = e^{-\lambda t^k}$$

From this formula, we know that the survival probability decreases more slowly over time than in models with increasing hazards, indicating that it is quite common to have early failures.The graphs below show the trend of the decreasing hazard function and related survivor function.

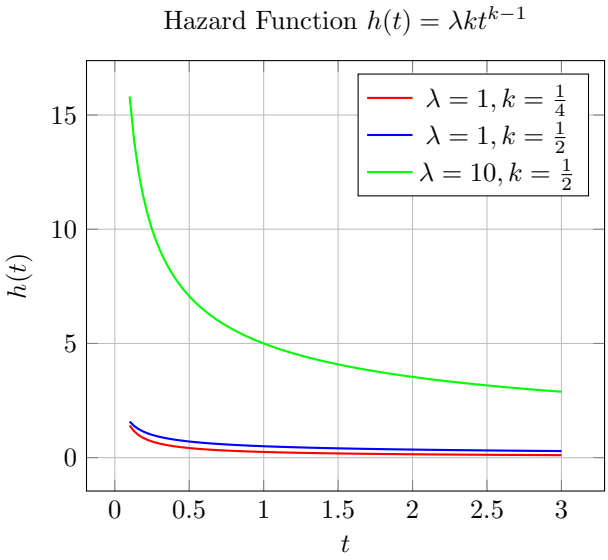


Figure 3: Hazard Function for Different λ and k

From the graph we find that:

- **When λ is fixed and k increases (e.g., from red to blue)**, the curve decreases more steeply at early time but also flattens out faster, reflecting a higher initial hazard but more rapid decay.
- **When k is fixed and λ increases (e.g., from blue to green)**, the entire hazard function moves upward, showing a uniformly higher failure rate at all time points.
- **When both λ and k increase**, the hazard function starts at a higher value and remains elevated for a longer period.

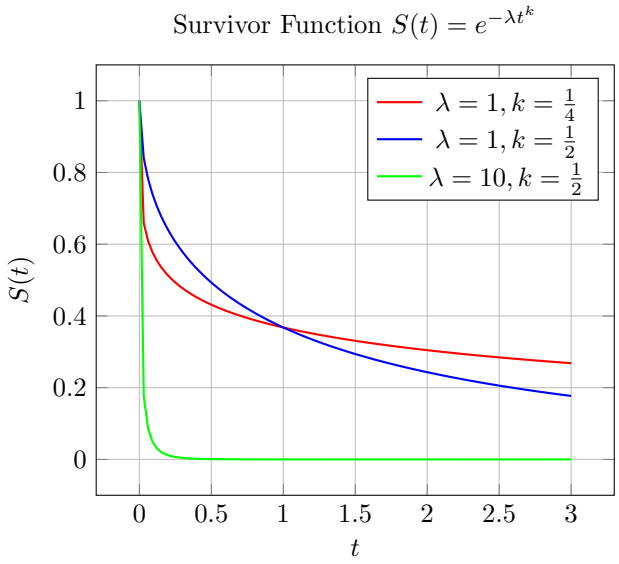


Figure 4: Survivor Function for Different λ and k

From the graph we find that the survivor function $S(t)$ decreases more rapidly as k increases indicating that larger k leads to a faster decline in survival probability over time. Additionally, when both k and λ increase, the survivor function decreases much more rapidly, meaning a significantly shorter survival time.

For the Gompertz distribution, the corresponding hazard function is

$$h(t) = \lambda e^{\beta t}$$

where:

- $\lambda > 0$ is the baseline hazard rate.
- $\beta > 0$ is the rate of aging or rate of increase in hazard over time.

From the formula, we can see that the hazard function exhibits exponential growth, compared to the polynomial growth observed in the Weibull model, which increases more rapidly over time. From $h(t) = \lambda e^{\beta t}$, we can get the survivor function:

$$S(t) = e^{-\frac{\lambda}{\beta}(e^{\beta t}-1)}$$

The graphs below show the trend of the Gompertz hazard function and related survivor function.

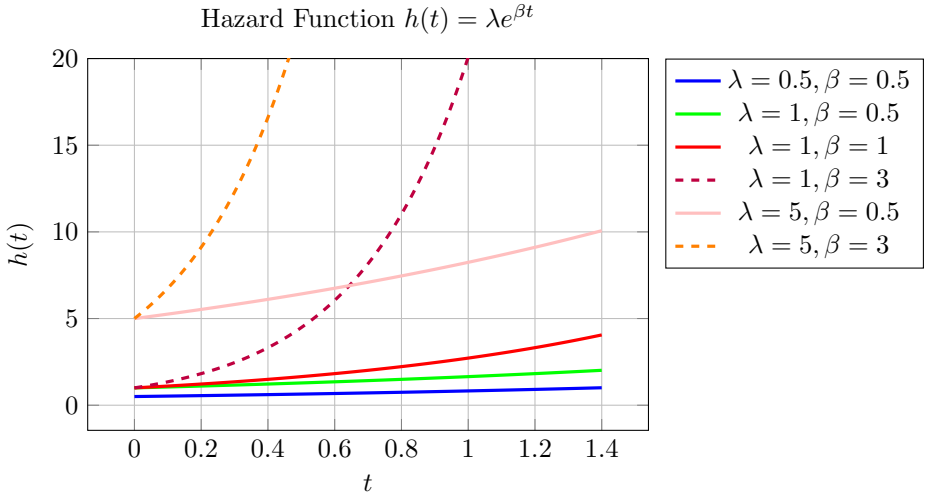
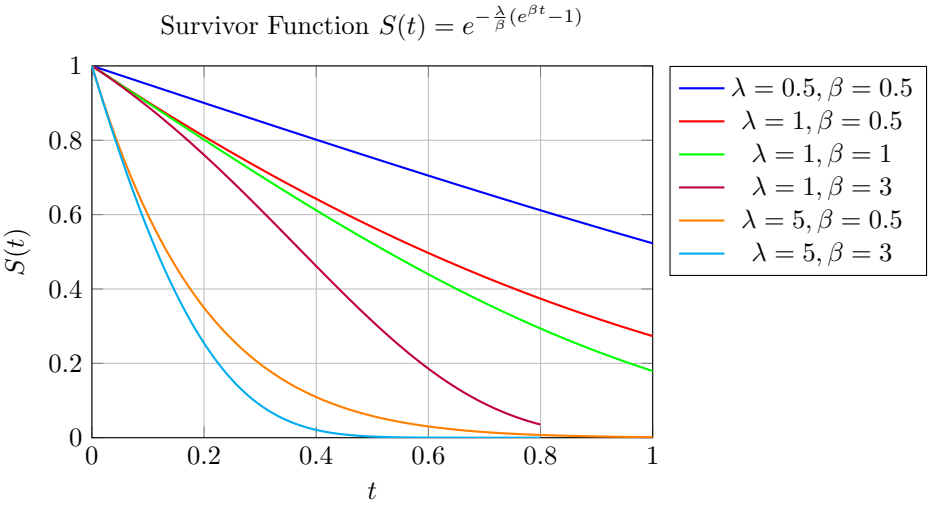


Figure 5: Hazard function for different λ and β values

From this graph, we find that λ and β determine the different characteristics of the curves. The hazard function always increases since β is larger than 0. As β increases, the rate at which the failure rate rises over time also accelerates. The parameter λ indicates the intercept of the curve; as it increases, it can also contribute to accelerating the rate of increase in the failure rate.



From the graph, we can find that a higher β leads to faster decay of $S(t)$, meaning that survival time is shorter and a higher λ also reduces survival probability, meaning increased hazard (faster failure). Furthermore, when both λ and β are large, the survival probability drops extremely fast. Additionally, when both λ and β are small, the survival probability remains high for a longer period of time.

Furthermore, a special case arises when $k = 1$ in the Weibull distribution or $\beta = 0$ in the Gompertz distribution.
Under this circumstance, the hazard function becomes constant. Then the function becomes:

$$h(t) = \lambda, \ \textit{where } \lambda > 0$$

which means the event occurs at a constant rate over time and is the characteristic of an exponential survival model. From $h(t) = \lambda$ we can get the survivor function:

$$S(t) = e^{-\lambda t}$$

which corresponds to an exponential distribution: $T \sim \text{Exp}(\lambda_0)$. So, the probability of surviving beyond t decreases exponentially. The graphs below show the trend of the constant hazard function and the related survivor function.

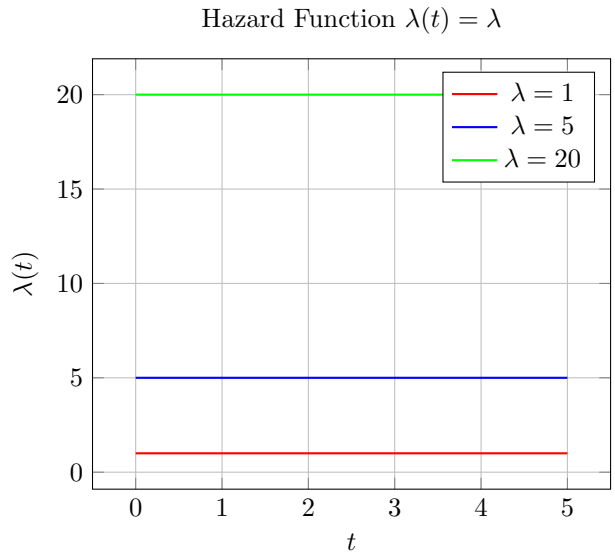


Figure 6: Hazard Function for Different λ

From the graph, we find that the hazard function is flat, which is independent of time. We can also draw the survivor function:

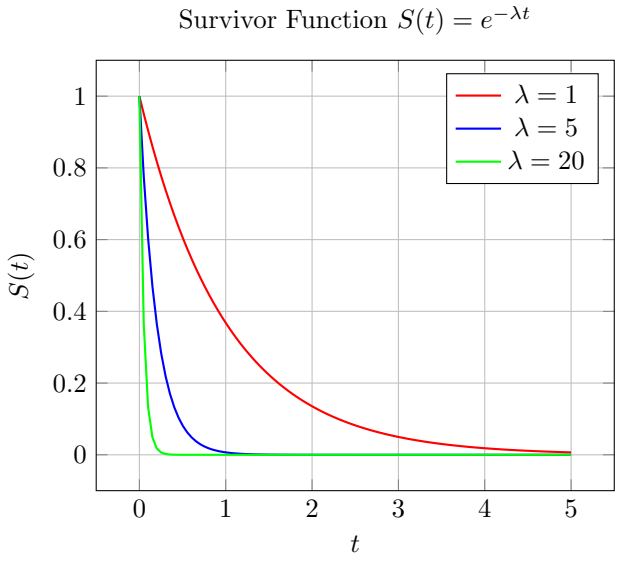


Figure 7: Survivor Function for Different λ

Since the survivor function: $S(t) = e^{-\lambda t}$, we substitute λ with 1, 5, and 20, and we find that all curves show a decreasing exponential trend: We find that as λ increases, the curve moves closer to the origin.

So, to briefly sum up, we can have the following table:

Table 1: Summary Table

Hazard Function	**Formula**	**Survivor Function**
Weibull (Increasing: $k > 1$, decreasing: $k < 1$)	$\lambda k t^{k-1}$	$e^{-\lambda t^k}$
Gompertz	$\lambda e^{\beta t}$	$e^{-\frac{\lambda}{\beta}(e^{\beta t}-1)}$
Constant ($k = 1$ for Weibull, $\beta = 0$ for Gompertz)	λ	$e^{-\lambda t}$

4 Censored data

4.1 Definition

Censoring in Survival Analysis refers to situations where the exact time of an event of interest (such as death, disease recurrence, or failure) cannot be fully observed for all subjects. It is typically assumed that censoring is non-informative, which means that the reason for censoring is independent of the subject's actual event time ensuring that the censored observations do not bias the estimation of survival probabilities. [4]
The main reasons for this are :

- The individuals are lost to follow up
- The study ends but no event occurs
- The participants withdraw from the study

There are three types of censoring:

- Right Censoring: The event occurs after the end of the study, but the exact time is unknown.
- Left Censoring: The event occurs before the study begins, but the exact time is unknown.
- Interval Censoring: The event occurs within a certain time interval, but the exact time is unknown.

For right censoring: we can use Kaplan-Meier estimator and Cox proportional hazards model to handle the data.

For left censoring: Accelerated Failure Time (AFT) models and Weighted Cox proportional hazards models are often used to cope with the data.

For interval censoring: Parametric survival models tend to be useful to deal with the data.

5 Kaplan-Meier Survival Curve

5.1 Definition of Kaplan-Meier Curve and its Confidence Interval

A Kaplan-Meier survival curve is a non-parametric estimator of the survival function which is a step function showing estimated survival probabilities over time [4]. Rather than making any assumptions about the underlying distribution of survival times, it directly constructs the estimate from observed data. Moreover, it also takes censored observations into account [4].

The Kaplan-Meier estimator for the survival function is given by:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where:

- t_i are the observed event times.

- d_i is the number of events (failures) at time t_i.
- n_i is the number of individuals at risk just before time t_i.

This formula represents the probability that an individual survives beyond time t by multiplying conditional survival probabilities at each event time.Here is an example of time to death (in weeks) for 10 patients who are diagnosed with lung cancer after treatment initiation:

Time t_i	**At Risk** n_i	**Events** d_i	**Censored**	$\hat{S}(t)$
0	10	0	0	1.00
3	10	1	0	0.90
5	9	1	0	0.80
8	8	1	0	0.70
10	7	1	0	0.60
13	6	1	0	0.50
15	5	1	0	0.40
18	4	1	0	0.30
20	3	1	0	0.20

Table 2: Kaplan–Meier Estimator Table with Risk Set and Event Counts

Then the graph is shown below:

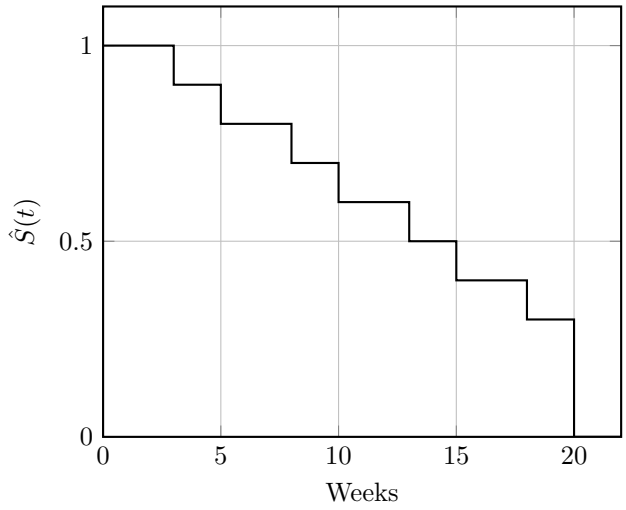


Figure 8: Kaplan-Meier Survival Curve

The survival probability $\hat{S}(t)$ decreases in a stepwise manner, rather than as a smooth curve. It always starts from 1 and drops only when an event occurs and remains constant at other times. The right end never increases, as the survival rate can only decrease or stay the same.

The confidence interval for a Kaplan-Meier survival curve indicates a range of plausible values for the actual survival probability at a specific time t, quantifying the uncertainty in the estimation of the survival function.
Since the $\hat{S}(t)$ is a product of random variables, it is hard to calculate its variance directly. So we firstly take the logarithm:

$$\log \hat{S}(t) = \sum_{t_i \leq t} \log \left(1 - \frac{d_i}{n_i}\right)$$

Since d_i/n_i is typically small, we apply the first-order Taylor expansion to $1 - d_i/n_i$:

$$\log\left(1 - \frac{d_i}{n_i}\right) \approx -\frac{d_i}{n_i}$$

Then,

$$\log \hat{S}(t) \approx -\sum_{t_i \le t} \frac{d_i}{n_i}$$

Thus, the variance of $\log \hat{S}(t)$ is approximately:

$$\mathrm{Var}\left(\log \hat{S}(t)\right) \approx \sum_{t_i \le t} \mathrm{Var}\left(\frac{d_i}{n_i}\right)$$

We introduce h_i which refers to the hazard probability at time t_i and suppose $d_i \sim \mathrm{Binomial}(n_i, h_i)$, we have:

$$\mathrm{Var}(d_i) = n_i h_i (1 - h_i)$$

Since h_i is small,

$$\mathrm{Var}(d_i) = n_i h_i (1 - h_i) \approx n_i h_i$$

Thus:

$$\mathrm{Var}\left(\frac{d_i}{n_i}\right) = \frac{1}{n_i^2}\mathrm{Var}(d_i) \approx \frac{h_i}{n_i}$$

Since d_i is the number of failures at time t_i, n_i is the number of individuals at risk just before time t_i, $\frac{d_i}{n_i}$ is an empirical estimation of the hazard probability h_i. Then we get:

$$\mathrm{Var}\left(\frac{d_i}{n_i}\right) \approx \frac{h_i}{n_i} \approx \frac{1}{n_i}\frac{d_i}{n_i} = \frac{d_i}{n_i^2}$$

For better accuracy, Greenwood [6] proposed a refined approximation:

$$\mathrm{Var}\left(\frac{d_i}{n_i}\right) \approx \frac{d_i}{n_i(n_i - d_i)}$$

Thus:

$$\mathrm{Var}\left(\log \hat{S}(t)\right) \approx \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}$$

We aim to compute the variance of $\hat{S}(t)$ by applying the Delta Method [8] to the transformation:

$$\hat{S}(t) = \exp\left(\log \hat{S}(t)\right)$$

Let $X = \log \hat{S}(t)$, then $\hat{S}(t) = e^X$. Using the Delta Method, the variance of a function $Y = f(X)$ is approximated by:

$$\mathrm{Var}(Y) \approx \left(f'(\mathbb{E}[X])\right)^2 \cdot \mathrm{Var}(X)$$

Since $f(X) = e^X$, we have $f'(X) = e^X$, so:

$$\mathrm{Var}(\hat{S}(t)) = \mathrm{Var}(\exp(\log \hat{S}(t))) \approx \left(\mathbb{E}[e^X]\right)^2 \cdot \mathrm{Var}(X) = \left(\mathbb{E}[\exp(\log \hat{S}(t))]\right)^2 \cdot \mathrm{Var}(\log \hat{S}(t))$$

Since $\hat{S}(t)$ is an unbiased estimator of the true survival function $S(t)$ when censoring is independent, it is obvious to get $E[\hat{S}(t)] \approx \hat{S}(t)$, then:

$$\mathrm{Var}\left(\hat{S}(t)\right) \approx \left(\hat{S}(t)\right)^2 \cdot \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}$$

Under the central limit theorem assumption, the 95% CI for the KM curve can be expressed as:

$$\hat{S}(t) \pm z_{\alpha/2} \cdot SE(\hat{S}(t)) = \hat{S}(t) \pm z_{0.05/2} \cdot \sqrt{\mathrm{Var}[\hat{S}(t)]} = \hat{S}(t) \pm 1.96\sqrt{\mathrm{Var}[\hat{S}(t)]}$$

This approximation follows the derivation in [5].
To obtain separate Kaplan–Meier survival curves for males and females using the `lung` dataset (as introduced in Section 1.1), we can use the R code below [10]:

```
library(survival)
lung <- survival::lung
lung
lung$status <- ifelse(lung$status == 2, 1, 0)
lung$sex <- factor(lung$sex, levels = c(1, 2), labels = c("Male", "Female"))
km_fit <- survfit(Surv(time, status) ~ sex, data = lung)
summary(km_fit)

Call: survfit(formula = Surv(time, status) ~ sex, data = lung)

                sex=Male
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   11    138       3   0.9783  0.0124       0.9542        1.000
   12    135       1   0.9710  0.0143       0.9434        0.999
   13    134       2   0.9565  0.0174       0.9231        0.991
   15    132       1   0.9493  0.0187       0.9134        0.987
   26    131       1   0.9420  0.0199       0.9038        0.982
   30    130       1   0.9348  0.0210       0.8945        0.977
  ... ...

                sex=Female
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5     90       1   0.9889  0.0110       0.9675        1.000
   60     89       1   0.9778  0.0155       0.9478        1.000
   61     88       1   0.9667  0.0189       0.9303        1.000
   62     87       1   0.9556  0.0217       0.9139        0.999
   79     86       1   0.9444  0.0241       0.8983        0.993
   81     85       1   0.9333  0.0263       0.8832        0.986
  ... ...
```

From the survfit output, we can obtain the survival probability, number at risk, number of events, standard error, and 95% confidence intervals for males, which can be compared with the female group to analyze gender differences in survival.

5.2 Log-Rank Test

The Log-rank test is a non-parametric statistical method used in survival analysis to determine if there is a significant difference between two or more survival curves. It is mainly utilized to assess whether the survival distributions of different treatment groups are the same [4].
In order to perform the log-rank test, we first state the null hypothesis: H0: There is no difference between survival curves.
Secondly, we calculate the expected cell counts:

$$e_{1f} = \left(\frac{n_{1f}}{n_{1f} + n_{2f}}\right) \times (m_{1f} + m_{2f})$$

$$e_{2f} = \left(\frac{n_{2f}}{n_{1f} + n_{2f}}\right) \times (m_{1f} + m_{2f})$$

where

- n_{if} represents the number in the risk set i at failure time f.
- m_{if} represents the number of failures in group i at failure time f.

Thirdly, we use:

$$O_i - E_i = \sum_{f=1}^{numberoffailuretimes} (m_{if} - e_{if}) \quad i = 1, 2$$

to get the difference between observed and expected data.
Then we calculate its variance:

$$\mathrm{Var}(O_i - E_i) = \sum_j \frac{n_{1f} n_{2f} (m_{1f} + m_{2f})(n_{1f} + n_{2f} - m_{1f} - m_{2f})}{(n_{1f} + n_{2f})^2 (n_{1f} + n_{2f} - 1)} \quad i = 1, 2$$

Finally, we get the Log-rank statistic:

$$\frac{(O_i - E_i)^2}{\mathrm{Var}(O_i - E_i)} \quad i = 1, 2$$

It is quite complicated to calculate the log-rank statistic by hand since the number of observed values is quite large, so with the help of R, we can get the log-rank statistic for the `lung` data [10].

```
logrank_test <- survdiff(Surv(time, status) ~ sex, data = lung)
logrank_test
Call:
survdiff(formula = Surv(time, status) ~ sex, data = lung)

             N Observed Expected (O-E)^2/E (O-E)^2/V
sex=Male   138      112     91.6      4.55      10.3
sex=Female  90       53     73.4      5.68      10.3

 Chisq= 10.3  on 1 degrees of freedom, p= 0.001

p_value <- 1 - pchisq(logrank_test$chisq, df = length(logrank_test$n) - 1)
formatted_p <- sprintf("%.6f", p_value)
cat("Log-rank test p-value:", formatted_p, "\n")

> Log-rank test p-value: 0.001311
```

From this, we can come to the conclusion that:
Since $p \approx 0.0013 < 0.01$, we can reject the null hypothesis, and there is strong evidence that the difference in survival between men and women is significant.
In addition, the observed number of deaths for men is 112, which is higher than the expected number which is 91.6, representing that men have a higher mortality rate. Meanwhile, the observed number of deaths for women is 53, which is less than the expected number, meaning that women experience longer survival times.

5.3 Draw the example K-M curve with 95% CI based on the `lung` data

Based on the `lung` data, we can draw the plot of the K-M curve with 95% CI with the code:

```
ggsurvplot(km_fit,
           data = lung,
           conf.int = TRUE,
           pval = TRUE,
           risk.table = TRUE,
           risk.table.height = 0.3,
           legend.labs = c("Male", "Female"),
           legend.title = "Sex",
           xlab = "Time (days)",
           ylab = "Survival Probability",
           ggtheme = theme_minimal(),
           palette = c("#E74C3C", "#3498DB"),
           size = 1.2,
           legend = c(0.8, 0.8)
)
```

Then we can have the graph:

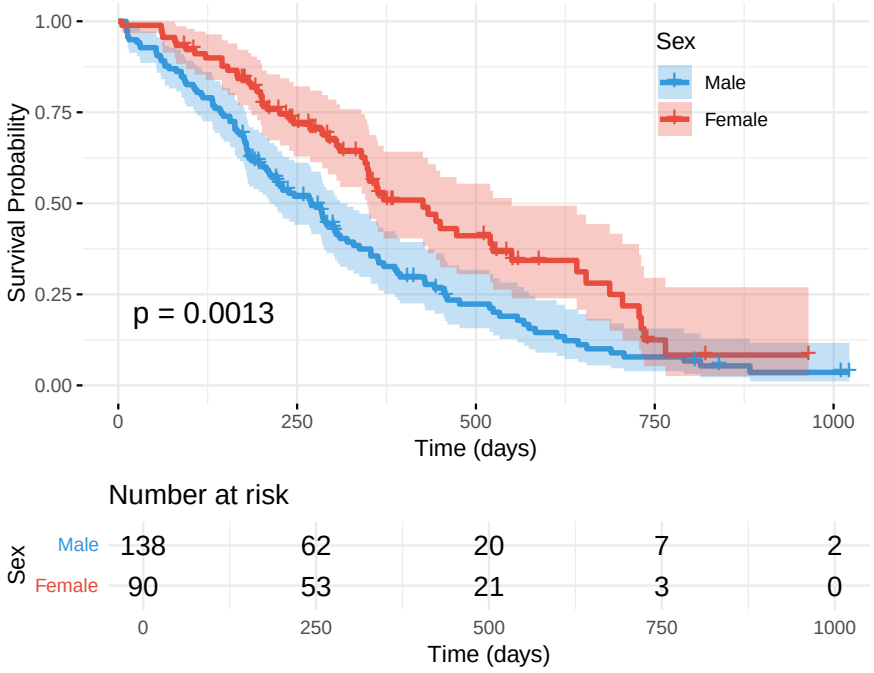


Figure 9: K-M curve with 95% CI based on the `lung` data

It is quite clear from the graph that:

- The Kaplan-Meier survival curve for male is below that of female, so it is obvious that the survival rate for female is higher than that for male.
- The 95% confidence interval creates a "region" around the curve, which gives a possible range of value that the specific time may obtain.
- The p-value approximately equals 0.0013, which conveys the same information as the plotted survival curves, indicating a significant difference between the groups.

6 Cox Proportional Hazards Model

6.1 Definition

The Cox proportional hazards model is a semiparametric survival analysis model that is widely used to analyze the relationship between survival time and one or more covariates [4]. The proportional hazards assumption is a key assumption of the Cox model, stating that the ratio of hazard functions between any two individuals remains constant over time.
Mathematically, the hazard function $h(t \mid \mathbf{x})$ which describes the instantaneous event rate at time t, given covariates $\mathbf{x}$, is as follows:

$$h(t \mid \mathbf{x}) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$$

where:
- $h(t \mid \mathbf{x})$ is the hazard function at time t for an individual with covariates $\mathbf{x}$.
- $h_0(t)$ is the baseline hazard function, which represents the hazard when all covariates are zero.
- $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ is the vector of covariates.
- $\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$represents the relative risk associated with the covariates.
- $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ is the vector of regression coefficients, which measure the effect of each covariate on the hazard rate.

6.2 Likelihood Function and its derivation

The joint probability density function treated as a function of the unknown parameters is called likelihood function. The likelihood function for independent random variables X_i, which follows a specific distribution with pdf: $f(x_i \mid \theta_i)$, indicating that the random variables are continuous, can be expressed as:

$$L(\boldsymbol{\theta}; \mathbf{x}) = \prod_{i=1}^{n} f(x_i \mid \theta_i)$$

where $\mathbf{x} = (x_1, x_2, ..., x_n)$ is a set of observed data, $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_n)$ is the vector of unknown parameters associated with each observation. θ_i represents the *ith* term that may generate the i_{th} data x_i, $f(x_i \mid \theta_i)$ denotes the pdf of the random variable X. The log-likelihood function can be obtained by taking the logarithm of the likelihood function $L(\boldsymbol{\theta}; \mathbf{x})$, which is:

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \ln(L(\boldsymbol{\theta}; \mathbf{x})) = \sum_{i=1}^{n} \ln f(x_i \mid \theta_i)$$

For unknown parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_n)$ in variables X, the values $\hat{\boldsymbol{\theta}}$ that maximise the likelihood function are the maximum likelihood estimates (MLE) of $\boldsymbol{\theta}$, and can be found by taking the derivative of $\ell(\boldsymbol{\theta}; \mathbf{x})$ with the assumption that $\theta_i = \theta$, then we get:

$$\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{x})}{\partial \theta} = 0$$

By solving this equation we can find $\hat{\boldsymbol{\theta}}$, which is the MLE for $\boldsymbol{\theta}$.

6.3 Likelihood Ratio (LR Statistic)

Likelihood Ratio is obtained by calculating a ratio of two likelihood functions to decide which model is preferred [4].

In order to assess the effect of extra covariates, we define 2 models:

- **Null Model** (H_0): A simpler model (fewer covariates).
- **Full Model** (H_1): A more complex model (additional covariates).

For these two nested models,the likelihood ratio is defined as:

$$\Lambda = \frac{L_{\text{null}}}{L_{\text{full}}}$$

where:

- L_{null} is the likelihood of the null model.
- L_{full} is the likelihood of the full model.

We take the log transformation:

$$\ln \Lambda = \ln L_{\text{null}} - \ln L_{\text{full}}$$

To obtain a test statistic that follows a chi-square (χ^2_d) distribution, we multiply by -2:

$$-2 \ln \Lambda = -2(\ln L_{\text{null}} - \ln L_{\text{full}}) \sim \chi^2_d$$

The degrees of freedom equals to

$$d = \text{number of parameters in full model} - \text{number of parameters in null model}$$

So, the LR static is:

$$LR = -2(\ln L_{\text{null}} - \ln L_{\text{full}}) \sim \chi^2_d$$

The use of the LR statistic will be explained in the next section of the Cox-PH Model.

6.4 Maximal Likelihood Estimation of the Cox-PH Model

When it comes to Cox Proportional Hazards Model, we have:
The pdf of the hazard function is:

$$h(t \mid \mathbf{x}) = h_0(t)e^{\mathbf{x}\boldsymbol{\beta}}$$

However, the baseline hazard function $h_0(t)$ is unknown, so we cannot use it to calculate the likelihood function directly. Because of this, we use partial likelihood which focuses only on the covariates $\mathbf{x}$ to get the likelihood function instead [4].
At each event time t_i, define the risk set $R(t_i)$ as the set of individuals still at risk just before t_i. The probability that subject i fails in t_i, given that someone in $R(t_i)$ fails, is:

$$P(\text{subject } i \text{ fails} \mid R(t_i)) = \frac{h(t_i \mid x_i)}{\sum_{j \in R(t_i)} h(t_i \mid x_j)}$$

Substituting $h(t \mid \mathbf{x}) = h_0(t)e^{\mathbf{x}\boldsymbol{\beta}}$, we get:

$$P(\text{subject } i \text{ fails} \mid R(t_i)) = \frac{h_0(t_i)e^{x_i\beta}}{\sum_{j \in R(t_i)} h_0(t_i)e^{x_j\beta}}$$

The item $h_0(t)$ can be canceled out and we get:

$$P(\text{subject } i \text{ fails} \mid R(t_i)) = \frac{e^{x_i\boldsymbol{\beta}}}{\sum_{j \in R(t_i)} e^{x_j\boldsymbol{\beta}}}$$

We use this formula to calculate the likelihood function, which is:

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{e^{x_i\boldsymbol{\beta}}}{\sum_{j \in R(t_i)} e^{x_j\boldsymbol{\beta}}}$$

Where k is the total number of distinct observed failure times.

Then, taking the logarithm, we get:

$$\ell_p(\boldsymbol{\beta}) = \sum_{i=1}^{k} \left[x_i\boldsymbol{\beta} - \ln \sum_{j \in R(t_i)} e^{x_i\boldsymbol{\beta}} \right]$$

Finally, we get the partial likelihood function for the Cox Proportional Hazards Model. In order to find the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$, denoted as $\hat{\boldsymbol{\beta}}$, we just solve:

$$\frac{\partial \ell_p(\boldsymbol{\beta})}{\partial \beta} = 0$$

where $\beta_1 = \beta_2 = ... = \beta_k = \beta$.

And the LR for for the Cox Proportional Hazards Model is the same as that for Generalized Linear Model (GLM), which is:

$$LR = -2(\ln L_{\text{null}} - \ln L_{\text{full}})$$

When we know the likelihood function of the null and full model, the ratio can be obtained, which follows a chi-square ($\chi^2_{d,1-\alpha}$).

When only using LR to assess the model:
A small (large) LR value suggests very little (large) improvement in model fit after adding other terms.
When using chi-square (χ^2) to assess the model:
The critical value from a chi-square table for $d = x$ at $\alpha = 0.05$ (95% confidence level) is $\chi^2_{x,0.05}$, and compare it with the LR value,

- If $LR > \chi^2_{x,0.05} \rightarrow$ Reject H_0 (new model is significant).
- If $LR \leq \chi^2_{x,0.05} \rightarrow$ Fail to reject H_0 (new model is not significant).

7 Parametric Survival Model

7.1 Definition

A parametric survival model is one in which the survival time is assumed to follow a known distribution. These models estimate survival and hazard functions using specific parameters, enabling more accurate predictions about survival times [4].
Parametric survival models distinctly differ from non-parametric models, such as the Kaplan-Meier estimator, and semi-parametric models like the Cox proportional hazards model. They rely on explicit assumptions about the underlying survival time distribution.

7.2 Accelerated Failure Time Model and PH Model

The Accelerated Failure Time (AFT) assumption is a fundamental assumption in parametric survival models. It suggests that the effect of covariates is proportional with respect to the survival time.
Compared to AFT, the underlying assumption for the Proportional Hazard (PH) Model is that the effect of covariates is multiplicative with respect to the hazard [4].
We then use an example to illustrate the function that the AFT and PH models will have on the test groups: There are two groups:

- **Group A**: Patients receive Standard Therapy.
- **Group B**: Patients receive New Experimental Therapy.

The AFT model assumes that the new therapy scales survival time by a factor ψ, which is called the acceleration factor. Under this circumstance, the relationship between the survival functions of Group A and Group B is:

$$S_B(t) = S_A(\psi t), \quad t \geq 0, \quad \text{where } \psi \text{ is the acceleration factor.}$$

The PH model assumes that the treatment reduces the hazard rate of death by a constant proportion at all time points. Given this assumption, the relationship between the survival functions of Group A and Group B is: If $\psi = \exp(\beta)$, then:

$$S_B(t) = S_A([\exp(\beta)]t)$$

7.3 Akaike Information Criterion (AIC)

Akaike's Information Criterion (AIC) is a method for comparing the fit of models that are not nested by utilizing the $-2log(likelihood)$ statistic.

Increasing the number of parameters in a model will almost always make the fitted value closer to observed value, thus improve the log-likelihood. However, this may lead to overfitting. The AIC addresses this issue by penalizing models with more parameters, thereby promoting a balance between explanatory power and parsimony. In model comparison, the model with the lowest AIC value is generally preferred.

7.4 Common distributions used in parametric survival models

Regarding `lung` data and treating age as the only covariate, we apply 3 common distributions: Weibull distribution, Log-logistic distribution, and Gompertz distribution to AFT and PH models [12].
For AFT model, using R, we have the following code and outcomes related to these 3 distributions [11]:

- Weibull distribution

```
AFT_Weibull <- aftreg(Surv(start_time,lung$time,lung$status)~ sex + age,
               data = lung, dist = "weibull")
summary(AFT_Weibull)
```

```
Covariate            W.mean      Coef Time-Accn  se(Coef)    LR p
```

```
sex                                                          0.0019
            Male      0.562     0         1   (reference)
          Female      0.438    -0.382     0.683     0.127
age                  61.961     0.012     1.012     0.007   0.0739

Events                    165
Total time at risk         69593
Max. log. likelihood      -1147.1
LR test statistic         13.59
Degrees of freedom        2
Overall p-value           0.00111739
```

- Loglogistic distribution

```
AFT_Loglogistic <- aftreg(Surv(start_time,lung$time,lung$status)~ sex + age,
                   data = lung, dist = "loglogistic")
summary(AFT_Loglogistic)

Covariate            W.mean      Coef Time-Accn  se(Coef)    LR p
sex                                                          0.0007
            Male      0.562     0         1   (reference)
          Female      0.438    -0.478     0.620     0.140
age                  61.961     0.014     1.014     0.008   0.0661

Events                    165
Total time at risk         69593
Max. log. likelihood      -1152.9
LR test statistic         16.07
Degrees of freedom        2
Overall p-value           0.000324439
```

- Gompertz distribution

```
AFT_Gompertz <- aftreg(Surv(start_time,lung$time,lung$status)~ sex + age,
                data = lung, dist = "gompertz")
summary(AFT_Gompertz)

Covariate            W.mean      Coef Time-Accn  se(Coef)    LR p
sex                                                          0.0083
            Male      0.562     0         1   (reference)
          Female      0.438    -0.321     0.726     0.133
age                  61.961     0.011     1.011     0.007   0.0997

Events                    165
Total time at risk         69593
Max. log. likelihood      -1150.3
LR test statistic         10.12
Degrees of freedom        2
Overall p-value           0.00633901
```

For PH model, using R, we have the following code and outcomes related to these 3 distributions:

- Weibull distribution

```
ph_weibull <- phreg(Surv(start_time,lung$time,lung$status)~ sex + age,
              data = lung, dist = "weibull")
summary(ph_weibull)

Covariate             Mean       Coef     Rel.Risk   S.E.    LR p
```

```
sex                                                          0.0019
            Male      0.562     0         1 (reference)
          Female      0.438    -0.507     0.602     0.167
age                  61.961     0.016     1.016     0.009   0.0739

Events                    165
Total time at risk         69593
Max. log. likelihood      -1147.1
LR test statistic         13.59
Degrees of freedom        2
Overall p-value           0.00111739
```

- Loglogistic distribution

```
ph_Loglogistic <- phreg(Surv(start_time,lung$time,lung$status)~ sex + age,
                  data = lung, dist = "loglogistic")
ph_Loglogistic

Call:
phreg(formula = Surv(start_time, lung$time, lung$status) ~ sex +
    age, data = lung, dist = "loglogistic")

Covariate          W.mean      Coef Exp(Coef)  se(Coef)    Wald p
(Intercept)                  19.420           27191.330     0.999
sex
            Male    0.562     0         1           (reference)
          Female    0.438    -0.507     0.602     0.167     0.002
age                 61.961     0.016     1.016     0.009     0.077

log(scale)                   21.301           20504.112     0.999
log(shape)                    0.282               0.069     0.000

Events                    165
Total time at risk         69593
Max. log. likelihood      -1147.1
LR test statistic         13.60
Degrees of freedom        2
Overall p-value           0.0011112
```

- Gompertz distribution

```
ph_Gompertz <- phreg(Surv(start_time,lung$time,lung$status)~ sex + age,
                  data = lung, dist = "gompertz")
summary(ph_Gompertz)

Covariate             Mean       Coef     Rel.Risk   S.E.    LR p
sex                                                          0.0027
            Male      0.562     0         1 (reference)
          Female      0.438    -0.489     0.613     0.167
age                  61.961     0.016     1.016     0.009   0.0747

Events                    165
Total time at risk         69593
Max. log. likelihood      -1148.9
LR test statistic         12.90
Degrees of freedom        2
Overall p-value           0.00157781
```

Across all models, the p-values for sex are less than 0.01, suggesting it to be a statistically significant predictor of survival. This indicates that there is a significant difference in survival between men and

women. While age appears to be a marginally significant predictor of survival, as its p-values remain between 0.05 and 0.1, which means that age may have an effect on the hazard.

7.5 Frequentist approach to the `lung` data

We then fit a Weibull accelerated failure time (AFT) model [17] and consider age, sex, wt.loss, meal.cal, phkarno_scaled and patkarno_scaled as five potential covariates that may have an effect on the survival time.

```
start_time <- rep(0, nrow(lung_eff))
AFT_Test_Weibull <- aftreg(Surv(start_time,lung_eff$time.m,lung_eff$status)~ 1 + age
                           + sex + wt.loss + meal.cal + phkarno_scaled
                           + patkarno_scaled, data = df, dist = "weibull")
summary(AFT_Test_Weibull)
```

```
> summary(AFT_Test_Weibull)
Covariate            W.mean      Coef Time-Accn  se(Coef)    LR p
age                  62.127     0.006     1.006     0.008   0.4686
sex                                                         0.0151
           Male       0.582     0         1   (reference)
         Female       0.418    -0.358     0.699     0.146
wt.loss              10.082    -0.006     0.994     0.005   0.2524
meal.cal            948.369    -0.000     1.000     0.000   1.0000
phkarno_scaled        0.828     0.011     1.011     0.570   1.0000
patkarno_scaled       0.813    -1.423     0.241     0.550   0.0094

Events                    121
Total time at risk        1736.3
Max. log. likelihood      -427.83
LR test statistic         16.50
Degrees of freedom        6
Overall p-value           0.0113072
```

From the output, we find that the LR p-value for `sex` is 0.0151, which is less than 0.05, indicating that there is evidence that sex will have an effect on survival time. Since males are set as the baseline, the coefficient of females is -0.358, which means that the failure rate for females is lower than that for males. So the survival time for females is longer than males. Besides, the LR p-value for `patkarno_scaled` is 0.0094, which is less than 0.01. So there is strong evidence that patkarno_scaled will influence the survival time. The coefficient is -1.423, which suggests that a higher score will, to some extent, decrease the failure rate, leading to a longer survival time. While the LR p-values for the rest of the covariates are larger than 0.1. So, there is no evidence that they will affect the survival time. Moreover, we can see that the overall p-value is approximately 0.011, which is less than 0.05, indicating that it is better than no model but that is not necessary a good model.

8 Bayesian Statistics

8.1 Brief Introduction

Bayesian statistics is a framework for statistical inference in which probabilities are used to quantify uncertainty about unknown parameters of interest. It is based on Bayes' Theorem, which updates prior beliefs about parameters or models using observed data to obtain a posterior distribution [13].

8.2 Bayes' Theorem

Bayes' Theorem mathematically expresses how prior beliefs about a parameter should be updated after observing new data:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{\int p(\mathcal{D} \mid \theta)\, p(\theta)\, d\theta} \tag{1}$$

Where:

- θ is the unknown parameter of interest,
- $\mathcal{D}$ denotes the observed data,
- $p(\theta)$ is the prior distribution,
- $p(\mathcal{D} \mid \theta)$ is the likelihood function (discussed in the previous section),
- $p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta)\, p(\theta)\, d\theta$ is the marginal likelihood,
- $p(\theta \mid \mathcal{D})$ is the posterior distribution.

- Prior distributions:
The prior distribution $p(\theta)$ represents the beliefs or knowledge about the parameter θ before any data are observed.

- Likelihood distribution:
The likelihood function $p(\mathcal{D} \mid \theta)$ represents the probability of observing the data $\mathcal{D}$, given a specific value of the parameter θ. It reflects how plausible different parameter values are in light of the observed data.

- Posterior distribution:
The posterior distribution $p(\theta \mid \mathcal{D})$ is the updated belief about θ after incorporating the observed data. By using Bayes' Theorem, it combines the prior and the likelihood:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{p(\mathcal{D})} \propto p(\mathcal{D} \mid \theta)\, p(\theta) \tag{2}$$

Since: $p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta)\, p(\theta)\, d\theta$ tends to be a constant, so the signal "$\propto$" is used here.

8.3 Conjugate Prior

When the distribution for $p(\theta)$ is known, after combining with the likelihood function $p(\mathcal{D} \mid \theta)$ of a particular model, the resulting posterior $p(\theta \mid \mathcal{D})$ follows the same distribution as $p(\theta)$.

The posterior distribution resulting from conjugate priors can be expressed analytically, without requiring numerical integration or sampling. In addition, conjugate priors can lead to computational convenience.Taking Beta-Binomial Conjugate Prior as an example:

Suppose we observe x successes in n Bernoulli trials, modeled as:

$$x \mid \theta \sim \text{Binomial}(n, \theta)$$

Let the prior distribution be:

$$\theta \sim \text{Beta}(\alpha, \beta)$$

Then the posterior is given by:

$$\begin{aligned} p(\theta \mid x) &\propto p(x \mid \theta) \cdot p(\theta) \\ &\propto \theta^x (1-\theta)^{n-x} \cdot \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ &\propto \theta^{x+\alpha-1} (1-\theta)^{n-x+\beta-1} \end{aligned}$$

Thus, the posterior distribution is:

$$\theta \mid x \sim \text{Beta}(\alpha + x, \beta + n - x)$$

It can be shown that both θ and $\theta \mid x$ follow Beta distribution, and the calculation process is simplified.

In addition, some typical conjugate priors are shown below:

Likelihood Distribution	Conjugate Prior	Posterior Distribution
$X \sim \text{Binomial}(n, \theta)$	$\theta \sim \text{Beta}(\alpha, \beta)$	$\theta \mid x \sim \text{Beta}(\alpha + x, \beta + n - x)$
$X \sim \text{Poisson}(\lambda)$	$\lambda \sim \text{Gamma}(\alpha, \beta)$	$\lambda \mid x \sim \text{Gamma}(\alpha + \sum x_i, \beta + n)$
$X \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 known	$\mu \sim \mathcal{N}(\mu_0, \tau^2)$	$\mu \mid x \sim \mathcal{N}(\mu_n, \tau_n^2)$, $\tau_n^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}, \quad \mu_n = \tau_n^2 \left(\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\tau^2}\right)$
$X \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 unknown	$\sigma^2 \sim \text{Inv-Gamma}(\alpha, \beta)$	$\sigma^2 \mid x \sim \text{Inv-Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{S}{2}\right)$
$X \sim \text{Exponential}(\lambda)$	$\lambda \sim \text{Gamma}(\alpha, \beta)$	$\lambda \mid x \sim \text{Gamma}(\alpha + n, \beta + \sum x_i)$

Table 3: Conjugate Priors and Posterior Distributions: Formula Summary with $x = (x_1, \ldots, x_n)$.

8.4 Markov Chain Monte Carlo (MCMC) and The Metropolis–Hastings algorithm

Since the conjugate priors might not always be appropriate, the posterior distribution can not be obtained [13]. Thus, MCMC is introduced to solve the general cases [9]. Here is the basic idea of MCMC:
From Bayes' Theorem

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{\int p(\mathcal{D} \mid \theta)\, p(\theta)\, d\theta}$$

It is sometimes difficult to calculate the integral $\int p(\mathcal{D} \mid \theta)\, p(\theta)\, d\theta$, especially when the parameter θ is high-dimensional. Moreover, if the likelihood function $p(\mathcal{D} \mid \theta)$ is analytically intractable or highly non-linear, this integral may not have a closed-form solution. Under such circumstances, the posterior distribution $p(\theta \mid \mathcal{D})$ cannot be obtained analytically.

Since posterior distributions tend to be non-standard, it typically does not follow any known families of distributions like the Normal, Gamma, or Beta, from which direct sampling is straightforward. So, instead we often construct a Markov chain that evolves step by step and eventually converges to $p(\theta \mid \mathcal{D})$, which means constructing a Markov chain $\{\theta^{(t)}\}_{t=0}^{\infty}$ such that, as $t \to \infty$, the distribution of $\theta^{(t)}$ converges in distribution to the target posterior distribution, mathematically,

$$\theta^{(t)} \xrightarrow{d} p(\theta \mid \mathcal{D}).$$

When it comes to the Markov chain, the important step is to determine a condition under which there exists an invariant distribution $\pi(\theta)$ to make:

$$\pi(\phi) = \sum_{\theta} \pi(\theta)\, p(\phi \mid \theta), \ \textit{or} \ \ \pi(\phi) = \int \pi(\theta)\, p(\phi \mid \theta)\, d\theta$$

for discrete variables and continuous variables, respectively, and $p(\phi \mid \theta)$ is the transition probability. In MCMC, we want to construct a transition probability $p(\phi \mid \theta)$ such that $p(\theta \mid \mathcal{D})$ is a stationary distribution. The transition probability can be regarded as a sum of the probability of remaining in the same state and jumping to another state.
Then $p(\phi \mid \theta)$ can be written as:

$$p(\phi \mid \theta) = p^*(\phi \mid \theta) + r(\theta)\delta(\phi \mid \theta),$$

Thus:

$$1 = \int_{\phi} p(\phi \mid \theta) = \int_{\phi} p^*(\phi \mid \theta) + r(\theta)$$

Where:

- $p^*(\phi \mid \theta)$ is the probability of jumping to another state.
- $r(\theta)$ represents the probability for the chain to remain in the same state.

- $p^*(\theta \mid \theta) = 0$ if $\delta(\theta \mid \theta) = 1$ and $\delta(\phi \mid \theta) = 0$ if $\phi \neq \theta$.

For a Markov Chain to have a stationary distribution, it needs to be reversible.
Let $\pi(\theta)$ denote $p(\theta \mid D)$ for a Markov Chain, the probability $p^*(\phi \mid \theta)$ would satisfy the detailed balance which is an equation for reversibility:

$$\pi(\theta)\, p^*(\phi \mid \theta) = \pi(\phi)\, p^*(\theta \mid \phi),$$

Then we have

$$\begin{aligned}
\int_\theta \pi(\theta) p(\phi \mid \theta) d\theta &= \int_\theta \pi(\theta) p^*(\phi \mid \theta) d\theta + \int_\theta \pi(\theta) r(\theta) \delta(\phi \mid \theta) d\theta \\
&= \int_\theta \pi(\phi) p^*(\theta \mid \phi) d\theta + \pi(\phi) r(\phi) \\
&= \pi(\phi) \int_\theta p^*(\theta \mid \phi) d\theta + \pi(\phi) r(\phi) \\
&= \pi(\phi) \left\{1 - r(\phi)\right\} + \pi(\phi) r(\phi) \\
&= \pi(\phi).
\end{aligned}$$

which satisfies that:

$$\pi(\phi) = \int \pi(\theta)\, p(\phi \mid \theta)\, d\theta$$

So, $\pi(\phi) = p(\theta \mid \mathcal{D})$ is an invariant distribution and the Markov Chain constructed will converge to the posterior distribution $p(\theta \mid \mathcal{D})$.

The general transition probability can be obtained by the Metropolis-Hastings Algorithm which achieves this by combining a proposal distribution $q(\phi \mid \theta)$ with an acceptance probability $\alpha(\phi \mid \theta)$ [9] which makes the detailed balance equation:

$$\pi(\theta)\, q(\phi \mid \theta)\, \alpha(\phi \mid \theta) = \pi(\phi)\, q(\theta \mid \phi)$$

hold, and consequently, we get the acceptance rate

$$\alpha(\phi \mid \theta) = \min\left(1, \frac{\pi(\phi)\, q(\theta \mid \phi)}{\pi(\theta)\, q(\phi \mid \theta)}\right).$$

If the proposal distribution $q(\phi \mid \theta)$ is symmetric (i.e. $q(\phi \mid \theta) = q(\theta \mid \phi)$), the form can be simplified to

$$\alpha(\phi \mid \theta) = \min\left(1, \frac{\pi(\phi)}{\pi(\theta)}\right).$$

so that the probability of going from state θ to state ϕ is

$$p^*(\phi \mid \theta) = q(\phi \mid \theta)\alpha(\phi \mid \theta),$$

while the probability that the chain remains in its present state θ is

$$r(\theta) = 1 - \int_\phi q(\phi \mid \theta)\alpha(\phi \mid \theta) d\theta.$$

Then $p(\phi \mid \theta)$ can be written as:

$$p(\phi \mid \theta) = p^*(\phi \mid \theta) + r(\theta)\delta(\phi \mid \theta) = q(\phi \mid \theta)\alpha(\phi \mid \theta) + \left(1 - \int_\phi q(\phi \mid \theta)\alpha(\phi \mid \theta) d\theta\right) \delta(\phi \mid \theta).$$

Over time, the chain will produce samples from the posterior distribution and, therefore, the shape of the distribution, approximate expectation, confidence interval, and other quantities of interest can be estimated. This is the basic ideal of the Monte Carlo approximation as explained in the next paragraph.

Let θ be a parameter of interest and let $y_1, \ldots, y_n \in D$. Suppose we could sample N independent, random θ-values from the posterior distribution

$$p(\theta \mid y_1, \ldots, y_n): \quad \theta^{(1)}, \ldots, \theta^{(N)} \sim \text{i.i.d. } p(\theta \mid y_1, \ldots, y_n).$$

Then the empirical distribution of the samples $\{\theta^{(1)}, \ldots, \theta^{(N)}\}$ would approximate $p(\theta \mid y_1, \ldots, y_n)$, with the precision improving with larger N, which is known as a Monte Carlo approximation to $p(\theta \mid y_1, \ldots, y_n)$. Additionally, let $g(\theta)$ be any function and, by the law of large numbers, if $\theta^{(1)}, \ldots, \theta^{(N)}$ are i.i.d. samples from $p(\theta \mid y_1, \ldots, y_n)$, then

$$\frac{1}{N} \sum_{n=1}^{N} g(\theta^{(n)}) \to \mathbb{E}[g(\theta) \mid y_1, \ldots, y_n] = \int g(\theta) p(\theta \mid y_1, \ldots, y_n)\, d\theta \quad \text{as } N \to \infty.$$

Implying that as $N \to \infty$,

- $\bar{\theta} = \sum_{n=1}^{N} \theta^{(n)}/N \to \mathbb{E}[\theta \mid y_1, \ldots, y_n]$;
- $\sum_{n=1}^{N} (\theta^{(n)} - \bar{\theta})^2/(N-1) \to \text{Var}[\theta \mid y_1, \ldots, y_n]$

Since the Markov chain produces dependent observations, the samples generated from it are autocorrelated, and thus they are not i.i.d. However, the Monte Carlo can still be used. The Ergodic Theorem [7] states that for irreducible, aperiodic and positive recurrent Markov chain, we have:

$$\hat{\mu}_N = \frac{1}{N} \sum_{i=1}^{N} f(X^{(i)}) \xrightarrow{a.s.} \mathbb{E}_\pi[f(X)]$$

Although samples are dependent, we can still get the mean value but for variance, we encounter:

$$\text{Var}(\hat{\mu}_N) \approx \frac{\sigma^2_{\text{eff}}}{N}$$

where $\sigma^2_{\text{eff}} > \sigma^2$ represents the effective variance, which accounts for the correlation among the samples [16].

The following will give a simplified example of finding posterior distributions for age in `lung` data with the help of R.

In this example, we consider a simple linear regression model:

$$\text{time.m}_i = \beta_0 + \beta_1 \cdot \text{age}_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where we assume both β_0 and σ^2 are fixed for simplicity. Our goal is to infer the posterior distribution of the coefficient β_1 given the observed data. This approach ignores the censoring in the 'lung' data for the 'status' variable is not included in this model. While this is not suitable for real-world survival analysis, it serves as a pedagogical example of posterior computation using Bayesian inference.

Before we try to use the MCMC to estimate the posterior distribution for the coefficient of age, we try to specify the prior distribution and we use `lm` to fit the model.

```
X <- lung_eff$age
Y <- lung_eff$time.m
fit <- lm(Y ~ X)
```

Then we need the information about the coefficient of age, and we can easily do this by:

```
summary(fit)

Call:
lm(formula = Y ~ X)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-10.027  -4.752  -1.780   3.588  24.405

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.03312    3.70567   3.787 0.000213 ***
X           -0.05907    0.05856  -1.009 0.314622
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.96 on 166 degrees of freedom
Multiple R-squared:  0.006091, Adjusted R-squared:  0.0001037
F-statistic: 1.017 on 1 and 166 DF,  p-value: 0.3146
```

So we assume that the coefficient of age follows a normal distribution with mean = -0.06 and the standard error = 0.06. Then we use the code:

```
beta0_hat <- coef(fit)[1]
sigma_hat <- summary(fit)$sigma
```

to obtain the estimates of the intercept and residual standard deviation to simplify the likelihood function. Then we set the parameters for the prior distribution:

```
prior_mean <- -0.06
prior_sd <- 0.06
```

With all the preparations done, we try to use the MCMC with the Metropolis-Hastings Algorithm to find the posterior distribution [14]:

```
library(gridExtra)
n.sim <- 2100
beta1 <- rep(NA, n.sim)
beta1[1] <- -0.059

set.seed(123)
for (i in 2:n.sim) {
  # Proposal: Normal random walk
  beta1_new <- rnorm(1, mean = beta1[i - 1], sd = 0.005)

  # log-likelihood
  mu_new <- beta0_hat + beta1_new * X
  mu_old <- beta0_hat + beta1[i - 1] * X
  loglik_new <- sum(dnorm(Y, mean = mu_new, sd = sigma_hat, log = TRUE))
  loglik_old <- sum(dnorm(Y, mean = mu_old, sd = sigma_hat, log = TRUE))

  # log-prior
  logprior_new <- dnorm(beta1_new, mean = prior_mean, sd = prior_sd, log = TRUE)
  logprior_old <- dnorm(beta1[i - 1], mean = prior_mean, sd = prior_sd, log = TRUE)

  # MH acceptance ratio
  log_acc_ratio <- (loglik_new + logprior_new) - (loglik_old + logprior_old)

  if (log(runif(1)) < log_acc_ratio) {
    beta1[i] <- beta1_new
  } else {
    beta1[i] <- beta1[i - 1]
  }
}
```

```
beta_post<- beta1[seq(101, n.sim, by = 5)]

# Plotting
library(ggplot2)
library(gridExtra)

trace_plot <- ggplot(data.frame(iter = 1:n.sim, beta1 = beta1),
  aes(x = iter, y = beta1)) +
  geom_line(color = "darkgreen") +
  geom_vline(xintercept = 100, linetype = "dashed", color = "red") +
  labs(title = "Trace Plot of Beta_age",
       x = "Iteration",
       y = expression(beta[1])) +
  theme_minimal()

density_plot <- ggplot(data.frame(beta1 = beta_post), aes(x = beta1)) +
  geom_density(fill = "lightblue", alpha = 0.6) +
  stat_function(fun = function(x) dnorm(x, mean = prior_mean, sd = prior_sd),
                color = "darkblue", linetype = "dashed", size = 1.2) +
  geom_vline(xintercept = mean(beta_post), color = "red", linetype = "dotted") +
  labs(title = "Posterior Density of Beta_age",
       x = expression(beta[1]),
       y = expression(pi(beta[1]~"|"~y))) +
  theme_minimal()

grid.arrange(trace_plot, density_plot, ncol = 1)
```

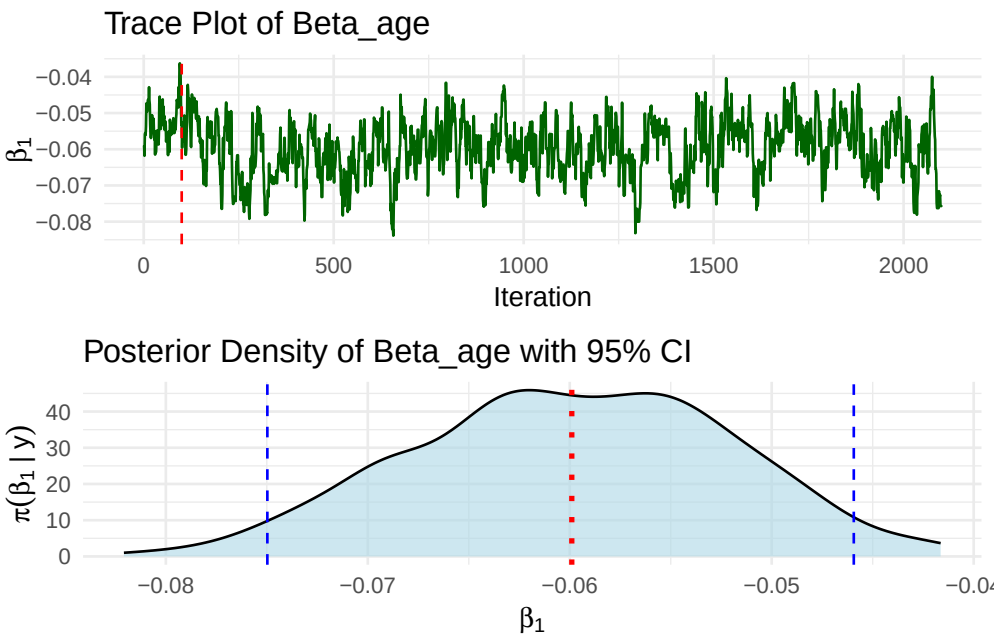


Figure 10: Posterior Distribution for scale parameter in distribution of age and its Iteration

The trace plot shows the sample values of β during about 2000 MCMC iterations. The chain appears to mix well, since there are rapid fluctuations between approximately -0.08 and -0.04 with some but not strong autocorrelation, which suggests a good exploration of the posterior distribution. Furthermore, no evident trend indicates that the chain is likely to be stable.

The density shape suggests that the posterior is approximately symmetric, which is centered around approximately-0.06, indicated by the red dashed line. We can also find that the 95% credible interval approximately lies between -0.075 and -0.046, which is below 0, suggesting a negative effect of age on survival time.

9 Bayesian Survival Analysis

Since Survival Analysis focuses on evaluating surviving time in clinical trials, and it is quite common to use the Cox proportional Hazard model, which is a non-parametric model. However, in practical implementation, the situation tends to be more complicated. Survival models are often difficult to fit, especially when faced with censoring. Bayesian statistics and MCMC techniques make the simulation more straightforward and, moreover, can make exact inference for any sample size. Additionally, it can naturally incorporate prior knowledge, account for parameter uncertainty, and provide full posterior distributions for quantities of interest, making it more available and flexible to model construction and data analysis [9].

9.1 Survival Analysis with Bayesian Statistics

Bayesian estimation

Let

- d represents a parametric family (e.g., exponential, Weibull, log-normal, etc.),
- $S_d(t)$ and $h_d(t)$ represent survival and hazard function of d respectively,
- t_i represent observed survival time for participant i,
- c_i denote the censoring indicator for participant i,
- β denotes the treatment effect of the dummy-coded treatment x_i,
- α_d denote the intercept,
- D denotes the observed data set.

So, the likelihood of the data {t, c} under a survival model M_d given parameters $\theta_d = (\beta, \alpha_d)$ is:

$$p(D \mid \theta_d, M_d) = \prod_{i=1}^{n} h_d(t_i \mid \mathbf{x}_i, \theta_d)^{I(c_i=1)} \cdot S_d(t_i \mid \mathbf{x}_i, \theta_d)$$

Assigning prior distributions $p(\theta_d \mid M_d)$ to each parameter to get the posterior distributions according to Bayes' Theorem:

$$p(\theta_d \mid D, M_d) = \frac{p(D \mid \theta_d, M_d) \cdot p(\theta_d \mid M_d)}{p(D \mid M_d)}$$

Where, $p(D \mid M_d)$ denotes the marginal likelihood, which is equal to $\int_{\theta_d} p(D \mid \theta_d, M_d) \cdot p(\theta_d \mid M_d)\, d\theta_d$ [15].

For `lung` data introduced in the first section of the article, we aim to analyze the relationship between survival time and five covariates: age(age), sex(sex), weight loss in the last six months (wt.loss), calories consumed at meals (meal.cal), Karnofsky performance score rated by physician (phkarno_scaled) and Karnofsky performance score rated by patient (patkarno_scaled).

We define the data frame first by

```
df <- data.frame(
  time         = lung_eff$time.m,
  status       = lung_eff$status,
  age          = lung_eff$age,
```

```
  sex          = lung_eff$sex,
  wt.loss      = lung_eff$wt.loss,
  meal.cal     = lung_eff$meal.cal,
  phkarno_scaled = lung_eff$ph.karno / 100,
  patkarno_scaled = lung_eff$pat.karno / 100
)
head(df)
```

Since we need to specify prior distributions for the intercepts and auxiliary parameters of the competing parametric families, we assume that we would expect the median survival time in the standard treatment group to be 9 months with an interquartile range of 8 months. In order to have a satisfactory degree of uncertainty about the parameter values, we set the standard deviation of the prior distributions to 0.5.

```
priors <- calibrate_quartiles(median_t = 9, iq_range_t = 8, prior_sd = 0.5)
priors
```

```
> priors
$intercept
$intercept$`exp-aft`
Normal(2.56, 0.5)
$intercept$`weibull-aft`
Normal(2.41, 0.5)
$intercept$`lnorm-aft`
Normal(2.2, 0.5)
$intercept$`llogis-aft`
Normal(2.2, 0.5)
$intercept$`gamma-aft`
Normal(1.38, 0.5)

$aux
$aux$`weibull-aft`
Lognormal(0.51, 0.28)
$aux$`lnorm-aft`
Lognormal(-0.69, 0.69)
$aux$`llogis-aft`
Lognormal(0.92, 0.19)
$aux$`gamma-aft`
Lognormal(0.94, 0.19)
```

Since we do not have strong prior knowledge about the effect of age, we use the parameter estimates obtained from the frequentist Weibull AFT model in section 7.5 as the basis for specifying prior distributions.

```
AFT_Test_Weibull <- aftreg(Surv(start_time,lung_eff$time.m,lung_eff$status)~ 1 + age
                                + sex + wt.loss + meal.cal + phkarno_scaled
                                + patkarno_scaled, data = df, dist = "weibull")
summary(AFT_Test_Weibull)
```

```
> summary(AFT_Test_Weibull)
Covariate            W.mean     Coef      Time-Accn  se(Coef)    LR p
age                  62.127     0.006     1.006      0.008       0.4686
sex                                                               0.0151
            Male      0.582     0         1   (reference)
          Female      0.418    -0.358     0.699      0.146
```

Specifically, we take the frequentist estimate of the age coefficient (0.006) as the prior mean, and to avoid overly informative priors, we set the prior standard deviation to a more diffuse value like 0.05, which is larger than the standard error (0.008) from the frequentist fit. So, we suppose $\theta_{\text{age}} \sim$

$\mathcal{N}(0.006, 0.05^2)$. Additionally, for the factor sex, we use `contrast = "treatment"` to set equal prior distribution on differences between the individual factor levels and the comparison level. We also do the same as age, and we suppose that $\theta_{\text{sex}} \sim \mathcal{N}(-0.4, 0.15^2)$. We also set the prior distributions for the wt.loss and meal.cal to be the t-student distribution and use the default parameters: degrees of freedom = 3, location = 0, scale = 1. Furthermore, we adjust for both of the Karnofsky performance scores by setting a wider centered standard normal prior distribution.

We then use the package RoBSA [18] to find the posterior distribution for age:

```
fit.est <- RoBSA(
  formula = Surv(time, status) ~ 1 + age + sex + wt.loss + meal.cal + phkarno_scaled
  + patkarno_scaled,
  data = df,
  priors = list(
    age              = prior("normal", parameters = list(mean = 0.006, sd = 0.05)),
    wt.loss          = prior("student", parameters = list(df = 3, location = 0,
                       scale = 1)),
    meal.cal         = prior("student", parameters = list(df = 3, location = 0,
                       scale = 1)),
    phkarno_scaled   = prior("normal", parameters = list(mean = 0, sd = 1)),
    patkarno_scaled  = prior("normal", parameters = list(mean = 0, sd = 1)),
    sex              = prior_factor("normal",
                                    parameters = list(mean = -0.4, sd = 0.15),
                                    truncation = list(0, Inf), contrast = "treatment")
  ),
  test_predictors = "",
  prior_intercept = priors[["intercept"]],
  prior_aux       = priors[["aux"]],
  seed = 123,
  rescale_data = TRUE,
  parallel = TRUE
)
summary(fit.est)

Call:
RoBSA(formula = Surv(time, status) ~ 1 + age + sex + wt.loss + meal.cal +
    phkarno_scaled + patkarno_scaled, data = df,
    priors = list(age = prior("normal",
                        parameters = list(mean = 0.006, sd = 0.05)),
    wt.loss = prior("student",
                        parameters = list(df = 3, location = 0, scale = 1)),
    meal.cal = prior("student",
                        parameters = list(df = 3, location = 0, scale = 1)),
    phkarno_scaled  = prior("normal", parameters = list(mean = 0, sd = 1)),
    patkarno_scaled = prior("normal", parameters = list(mean = 0, sd = 1)),
    sex = prior_factor("normal",
                        parameters = list(mean = -0.4, sd = 0.15),
                        truncation = list(0, Inf), contrast = "treatment")),
    test_predictors = "",
    prior_intercept = priors[["intercept"]],
    prior_aux = priors[["aux"]],
    parallel = TRUE,
    seed = 123,
    rescale_data = TRUE)

Robust Bayesian survival analysis
Distributions summary:
```

```
            Models   Prior prob.  Post. prob. Inclusion BF
exp-aft     2/10     0.200        0.005       0.019
weibull-aft 2/10     0.200        0.880       29.427
lnorm-aft   2/10     0.200        0.000       0.000
llogis-aft  2/10     0.200        0.048       0.201
gamma-aft   2/10     0.200        0.067       0.288

Model-averaged estimates:

                 Mean      Median     0.025      0.975
age              -0.017    -0.017     -0.099     0.066
sex[Female]      0.084     0.071      0.004      0.229
wt.loss          0.070     0.069      -0.068     0.211
meal.cal         -0.005    -0.007     -0.144     0.148
phkarno_scaled   0.019     0.016      -0.132     0.185
patkarno_scaled  0.218     0.219      0.051      0.381
```

From the second table, we can obtain a preliminary overview of the posterior distribution of the log-acceleration factor (log(AF)) associated with age when it is treated as a covariate in the survival model. We can see that the mean and median for age are both -0.017, suggesting a small negative effect. The 95% credit interval is [-0.099, 0.066] which contains 0, indicating that the effect of age on the survival time is uncertain.

We also find out that the meal.cal lead to shorter survival times with the mean model-averaged log(AF) = -0.005, 95% CI [-0.144, 0.148]. While, scaled patient Karnofsky performance score and sex(female) show a statistically significant positive effect on survival time. The former has a model-averaged log(AF) of 0.218 and a 95% credible interval [0.051, 0.381], and the log(AF) for the latter is 0.084, with a 95% credible interval [0.004, 0.229]. Other predictors, including weight loss and scaled physician Karnofsky score, have credible intervals that include zero, indicating insufficient evidence for an effect on survival. Additionally, we can draw the plot of the posterior distribution by using the code below:

```
diagnostics_density( fit.est , parameter = "age")
```

Then we can get

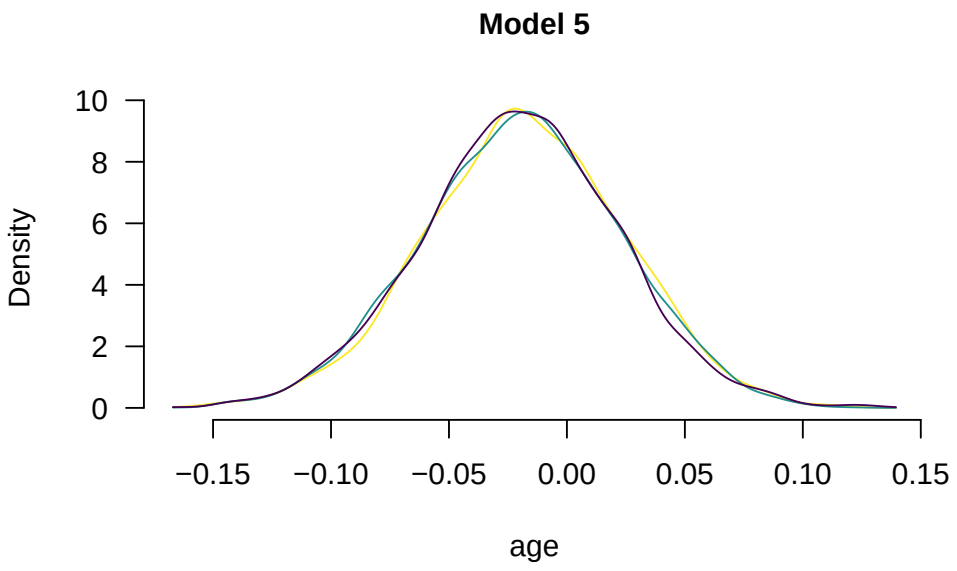


Figure 11: Posterior Distribution of the Age Scale Parameter in Bayesian Survival Analysis

In order to get posterior distributions for other covariates, just repeat the step.

```
diagnostics_density( fit.est , parameter = "sex")
diagnostics_density( fit.est , parameter = "wt.loss")
diagnostics_density( fit.est , parameter = "meal.cal")
diagnostics_density( fit.est , parameter = "phkarno_scaled")
diagnostics_density( fit.est , parameter = "patkarno_scaled")
```

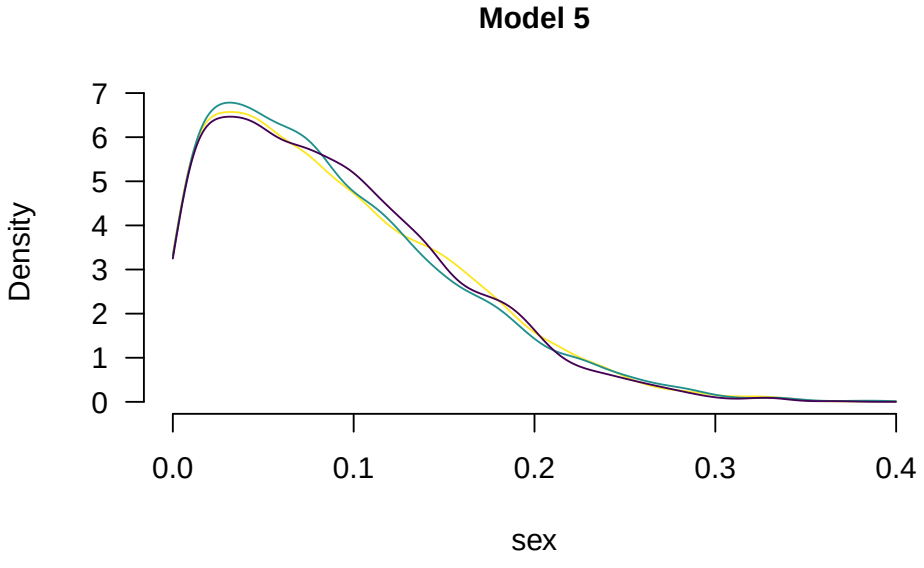


Figure 12: Posterior Distribution of the Sex Scale Parameter in the Bayesian Survival Analysis

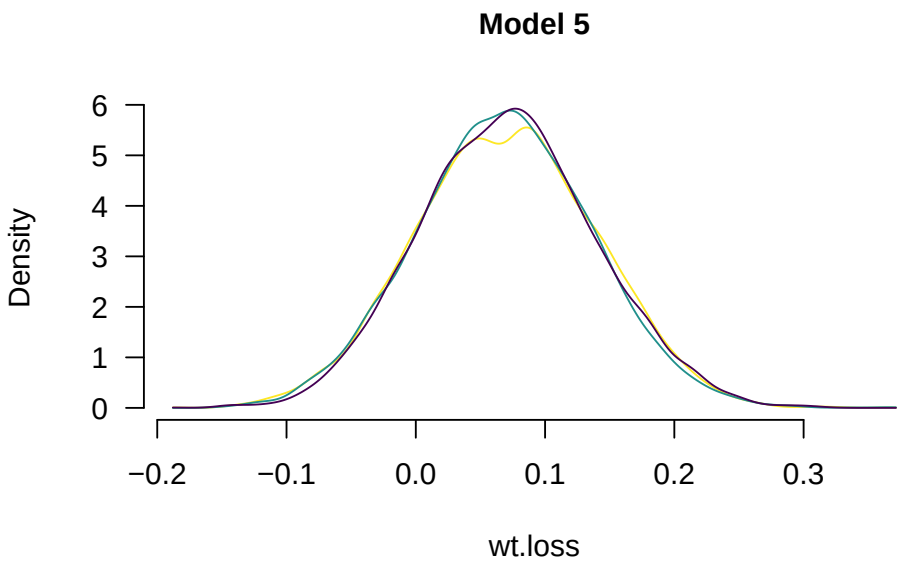


Figure 13: Posterior Distribution of the Wt.loss Scale Parameter in the Bayesian Survival Analysis

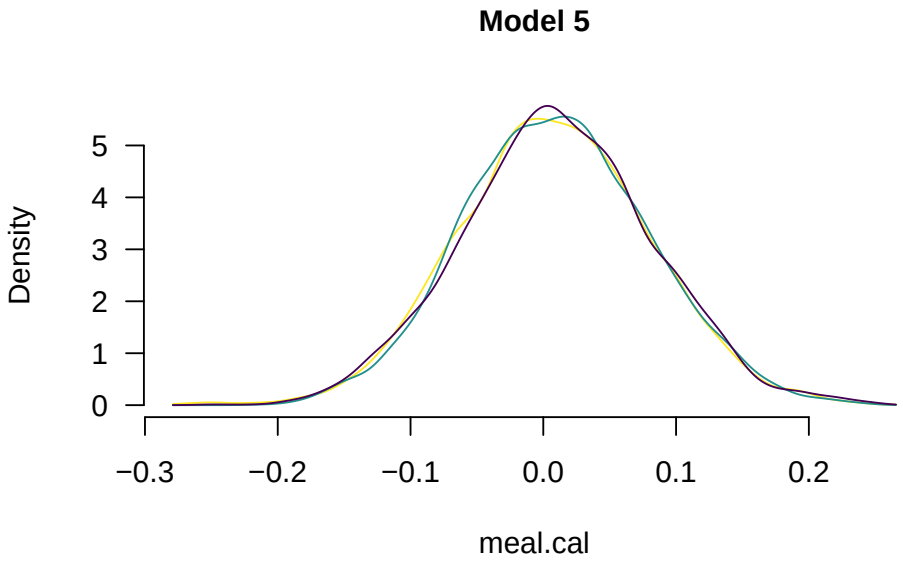


Figure 14: Posterior Distribution of the Meal.cal Scale Parameter in the Bayesian Survival Analysis

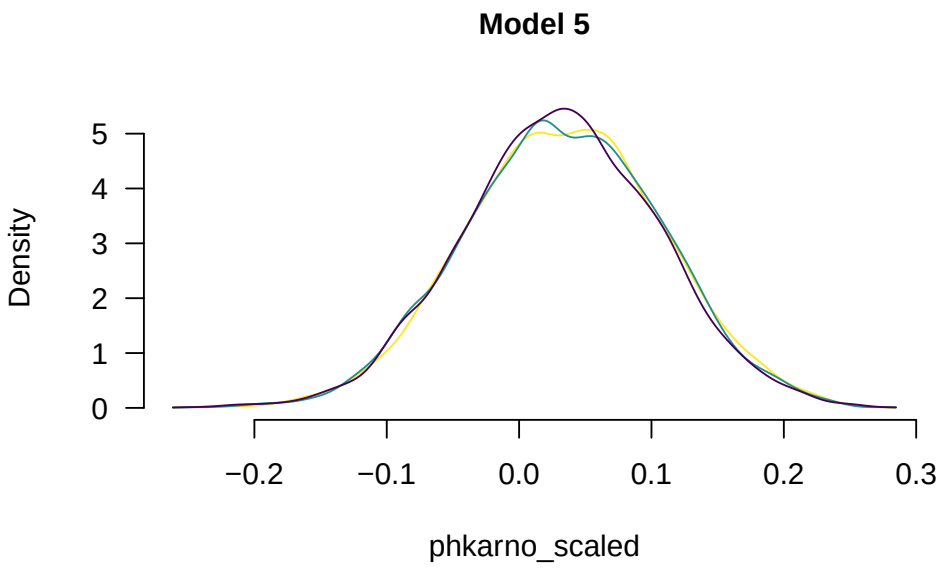


Figure 15: Posterior Distribution of the Phkarno scaled Parameter in the Bayesian Survival Analysis

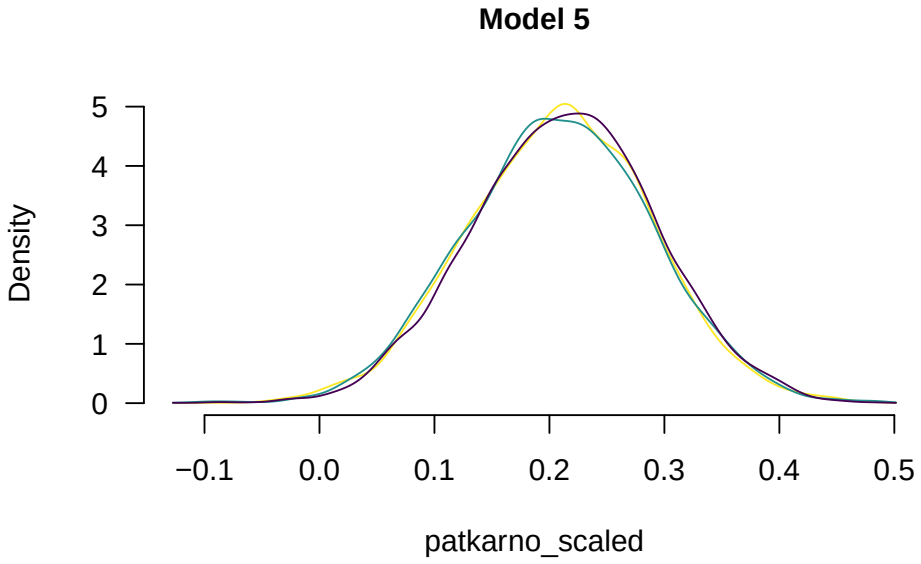


Figure 16: Posterior Distribution of the Patkarno scaled Parameter in the Bayesian Survival Analysis

Bayesian hypothesis testing

After obtaining the posterior distribution with the assumption that the treat has an effect, it should be compared to another model which assumes no effect of the treatment to get a quantified evidence [15].
We set $M_{0,d}$ the model with no treatment effect (i.e. $\beta = 0$), $M_{1,d}$ the model with treatment effect (i.e. $\beta = f_\beta$) conditioning on the parametric family d. Then the posterior probabilities based on different

models can be obtained:

$$p(M_{0,d} \mid D, d) = \frac{p(D \mid M_{0,d}, d) \cdot p(M_{0,d} \mid d)}{p(D \mid d)}$$

$$p(M_{1,d} \mid D, d) = \frac{p(D \mid M_{1,d}, d) \cdot p(M_{1,d} \mid d)}{p(D \mid d)}$$

where $p(D \mid d)$ can be obtained from the law of total probability:

$$p(D \mid d) = p(D \mid M_{0,d}, d) \cdot p(M_{0,d} \mid d) + D \mid M_{1,d}, d) \cdot p(M_{1,d} \mid d)$$

So, the whole process can be:
· **Step 1**: Set the null hypothesis H_0: the treatment has no effect and the alternative hypothesis H_1: the treatment has effect.
· **Step 2**: Calculate the Bayes Factor (BF):

$$\mathrm{BF}_{10} = \frac{p(D \mid M_1)}{p(D \mid M_0)}$$

It is a ratio of marginal likelihood, and the outcome can be interpreted as the support obtained for one model over the other.
· **Step 3** : Interpretation

- If $BF_{10} > 1$, Model 1 is preferred, then H_1 is preferred, with larger values providing increasingly stronger support for Model 1.
- If $BF_{10} < 1$, Model 0 is preferred, then H_0 is preferred, and smaller values correspond to stronger support for Model 0.
- If $BF_{10} = 1$, The evidence is indifferent between H_0 and H_1.

To assess the effect of treatment across all parametric models and derive a general conclusion regarding model preference, Bayesian model-averaging is often used. It regards models of different parametric families assuming presence of the treatment effect as a whole by adding them together:

$$p(D \mid M_1) = \sum_d p(D \mid M_{1,d}) \cdot p(M_{1,d})$$

and

$$p(D \mid M_0) = \sum_d p(D \mid M_{0,d}) \cdot p(M_{0,d})$$

for all models with no treatment effect.
Combing the treatment effect β of different parametric models with their own posterior distributions, we can get the model-averaging posterior distribution of β:

$$p(\beta \mid D, M_1) = \sum_d p(\beta \mid M_{1,d}, D) \cdot p(M_{1,d} \mid D)$$

$$p(\beta \mid D, M_0) = \sum_d p(\beta \mid M_{0,d}, D) \cdot p(M_{0,d} \mid D)$$

and we can also get the posterior model-averaged survival and hazard functions:

$$S(t) = \sum_d S_{1,d}(t) \cdot p(M_{1,d} \mid D) \quad , \quad S(t) = \sum_d S_{0,d}(t) \cdot p(M_{0,d} \mid D)$$

$$h(t) = \sum_d h_{1,d}(t) \cdot p(M_{1,d} \mid D) \quad , \quad h(t) = \sum_d h_{0,d}(t) \cdot p(M_{0,d} \mid D)$$

In addition, we can get the Bayes Factor under this circumstance:

$$\mathrm{BF}_{10} = \frac{\sum_d p(\mathcal{M}_{1,d}|D)}{\sum_d p(\mathcal{M}_{0,d}|D)} \Big/ \frac{\sum_d p(\mathcal{M}_{1,d})}{\sum_d p(\mathcal{M}_{0,d})}$$

If we want to compare a specific parametric model, take exponential as an example, to the rest of the models to find the best fit, the Bayes Factor will also help:

$$\mathrm{BF}_{\mathrm{exp}} = \frac{\sum_{m=0}^{1} p(\mathcal{M}_{m,1}|D)}{\sum_{m=0}^{1}\sum_{d'} p(\mathcal{M}_{m,d'}|D)} \Bigg/ \frac{\sum_{m=0}^{1} p(\mathcal{M}_{m,1})}{\sum_{m=0}^{1}\sum_{d'} p(\mathcal{M}_{m,d'})}$$

Then, in the following, we try to use Bayesian Model-averaging to assess the effect of age, with the help of R. We proceed by specifying a RoBSA model to test an informed hypothesis of the presence of the age effect [18].

```
library(RoBSA)

fit.test <- RoBSA(
  formula = Surv(time, status) ~ 1 + age + sex + wt.loss + meal.cal + phkarno_scaled
  + patkarno_scaled,
  data = df,
  priors = list(
    age              = prior("normal", parameters = list(mean = 0.006, sd = 0.05)),
    wt.loss          = prior("student",
                             parameters = list(df = 3, location = 0, scale = 1)),
    meal.cal         = prior("student",
                             parameters = list(df = 3, location = 0, scale = 1)),
    phkarno_scaled   = prior("normal", parameters = list(mean = 0, sd = 1)),
    patkarno_scaled  = prior("normal", parameters = list(mean = 0, sd = 1)),
    sex              = prior_factor("normal",
                                    parameters = list(mean = -0.4, sd = 0.15),
                                    truncation = list(0, Inf),
                                    contrast = "treatment")
  ),
  test_predictors = "age",
  prior_intercept = priors[["intercept"]],
  prior_aux       = priors[["aux"]],
  seed = 123,
  rescale_data = TRUE,
  parallel = TRUE
)

summary(fit.test)

Call:
RoBSA(formula = Surv(time, status) ~ 1 + age + sex + wt.loss +
    meal.cal + phkarno_scaled + patkarno_scaled, data = df,
    priors = list(age = prior("normal",
                              parameters = list(mean = 0.006, sd = 0.05)),
    wt.loss  = prior("student",
               parameters = list(df = 3, location = 0, scale = 1)),
    meal.cal = prior("student",
               parameters = list(df = 3, location = 0, scale = 1)),
    phkarno_scaled  = prior("normal",
                            parameters = list(mean = 0, sd = 1)),
    patkarno_scaled = prior("normal",
                            parameters = list(mean = 0, sd = 1)),
    sex = prior_factor("normal",
                       parameters = list(mean = -0.4, sd = 0.15),
                       truncation = list(0, Inf),
                       contrast = "treatment")),
    test_predictors = "age",
```

```
    prior_intercept = priors[["intercept"]],
    prior_aux = priors[["aux"]],
    parallel = TRUE,
    seed = 123,
    rescale_data = TRUE)

Robust Bayesian survival analysis

Distributions summary:

            Models   Prior prob.  Post. prob. Inclusion BF
exp-aft     2/10     0.200        0.005       0.018
weibull-aft 2/10     0.200        0.880       29.442
lnorm-aft   2/10     0.200        0.000       0.000
llogis-aft  2/10     0.200        0.048       0.203
gamma-aft   2/10     0.200        0.067       0.286

Components summary:

            Models   Prior prob.  Post. prob. Inclusion BF
age         5/10     0.500        0.474       0.902

Model-averaged estimates:

                 Mean      Median     0.025      0.975
age              -0.008    0.000      -0.087     0.053
sex[Female]      0.084     0.072      0.003      0.232
wt.loss          0.070     0.069      -0.064     0.211
meal.cal         -0.003    -0.005     -0.142     0.150
phkarno_scaled   0.023     0.021      -0.127     0.188
patkarno_scaled  0.218     0.219      0.048      0.381
```

In the first table, we see that the most posterior model probability is retained by the Weibull (0.880) family, and the related Bayes factor is 29.442 which indicates a strong evidence that Weibull distribution is suggested, quantifying the change from prior to posterior model probabilities.

The second table then summarizes information about hypothesis tests of the model components. We find the Bayes factor to be 0.902, indicating weak support for the null hypothesis that age has no effect compared to our informed hypothesis of an effect, which agrees with the previous section.

10 Conclusion

This project provides a comprehensive exploration of survival analysis from a theoretical and applied perspective. In Section 2, we introduced the development of survival analysis, and Section 3 detailed the mathematical definitions and derivations of the survivor and hazard functions and showed their relationship. In Section 4, we discussed the concept of censoring and introduced the non-informative assumption. The Kaplan-Meier estimator was introduced in Section 5 by providing mathematical formulas, example plots, and methods to compare survival curves using the log-rank test. Section 6 focused on the Cox Proportional Hazards Model, including likelihood-based estimation techniques and model interpretation. In Section 7, we introduced the Accelerated Failure Time (AFT) model and the PH model, which uses the acceleration factor to scale the time. Moreover, some commonly used distributions were included to fit the model. Section 8 offered an overview of Bayesian statistics, including conjugate priors and the MCMC method. Finally, in Section 9, we applied Bayesian survival analysis using R to `lung` data to find the posterior distribution of the scaled parameter of five potential covariates. We then did a hypothesis test on age to find that age may not have an effect on the survival time.

To evaluate the effect of age on survival time, we used the Weibull AFT model and Bayesian statistics

to conduct analyses. However, the outputs from these two methods are different. The Weibull AFT model is a typical frequentist approach that treats parameters as fixed but unknown and relies solely on the data for inference. While the Bayesian approach incorporates prior distributions and produces a full posterior distribution over parameters. So, the Bayesian approach can provide a more precise expression of uncertainty, whereas the frequentist method yields point estimates and confidence intervals without probabilistic interpretation.

From the Weibull AFT model, the coefficient for age was estimated to be 0.006 and the standard error is about 0.008, with an LR p-value of 0.4686. Although the positive coefficient suggests a slight increase in survival time with age, the effect was not statistically significant ($0.4686 > 0.05$). While in the Bayesian analysis, we specified a weakly informative prior which is $\theta_{\text{age}} \sim \mathcal{N}(0.006, 0.05^2)$. The posterior mean of the coefficient for age was –0.017, with a 95% credible interval of [–0.099, 0.066]. As the credible interval contains zero, no definitive conclusion can be drawn regarding the direction between age and survival time. It is interesting to see that the frequentist approach suggests that age is positively related to survival time. While the outcome from the Bayesian approach reveals that survival time will decrease as age increases, which is consistent with the actual situation. So, it is often the case that the posterior distribution will give a more precise conclusion of the effect of the covariate with more information included by likelihood function.

In addition, there are restrictions on the analysis. The original data has too many "NA" terms, and in order to fit a proper model, it is required to omit the influence of these "NA" values. So, the data frame used in the Bayesian estimation and hypothesis testing only contains the adjusted values, which may lead to sample selection bias. Furthermore, the analysis is restricted to five covariates and five AFT models, which may limit the adequacy of the model. Besides, interaction should also be taken into account, but the related prior distributions are difficult to find.

A Code for Parametric Survival Models and Bayesian Survival Models

```
library(eha)
library(survival)
library(survminer)
lung <- survival::lung
lung$inst <- NULL
lung$sex <- as.factor(lung$sex)
levels(lung$sex) <- c("Male", "Female")
lung$status <- ifelse(lung$status == 2, 1, 0)
lung$time.m <- round(lung$time / 30, 3)
lung$time <- NULL
lung_eff <-na.omit(lung)
head(lung_eff)
```

References

[1] Loprinzi, C. L., *et al.* (1994). *Prospective evaluation of prognostic variables from patient-completed questionnaires. Journal of Clinical Oncology*, **12**(3), 601–607.

[2] Crooks, V., Waller, S., Smith, D., & Hahn, T. J. (1991). *The use of the Karnofsky Performance Scale in determining outcomes and risk in geriatric outpatients. Journal of Gerontology*, **46**(4), M139–M144.

[3] Liberato Camilleri. *History of Survival Analysis.* The Sunday Times of Malta, LIFE & WELLBEING section, March 24, 2019, p. 53.

[4] David G. Kleinbaum and Mitchel Klein (2012). *Survival Analysis: A Self-Learning Text*, Springer.

[5] Charles Zaiontz. *Real Statistics Resource Pack (Release 8.9.1)*. Copyright (2013–2023).

[6] Greenwood, M. (1926). *The natural duration of cancer.* Reports on Public Health and Medical Subjects, **33**, 1–26. Her Majesty's Stationery Office, London.

[7] Petersen, K. *Ergodic Theory.* Corrected reprint of the 1983 original. Cambridge Studies in Advanced Mathematics, Vol. 2. Cambridge University Press, Cambridge, 1989. xii + 329 pp.

[8] George Casella and Roger L. Berger. *Statistical Inference*, 2nd edition. Duxbury Press, 2002.

[9] Peter M. Lee (2012). *Bayesian Statistics: An Introduction*, John Wiley & Sons, Incorporated, Newark.

[10] Flore Uzan (2020). *Lung Data - Survival Analysis*, published on RPubs.

[11] Göran Broström (2021). *Event History Analysis with R, Second Edition.*

[12] Khan, S. and Khosa, S. (2016). *Generalized log-logistic proportional hazard model with applications in survival analysis. Journal of Statistical Distributions and Applications*, **3**(1), Article 16.

[13] Ibrahim, Joseph G., Chen, Ming-Hui, and Sinha, Debajyoti (2001). *Bayesian Survival Analysis.* Springer, New York, NY.

[14] Gómez-Rubio, Virgilio (2020). *Bayesian Inference with INLA*. Chapman & Hall/CRC Press, Boca Raton, FL.

[15] Bartoš, F., Aust, F. and Haaf, J.M. (2022). *Informed Bayesian survival analysis. BMC Medical Research Methodology*, **22**, Article 238.

[16] Hoff, Peter D. (2009). *A First Course in Bayesian Statistical Methods.* Springer, New York.

[17] Muse, A. H., *et al.* (2022). Bayesian and frequentist approach for the generalized log-logistic accelerated failure time model with applications to larynx-cancer patients. *Alexandria Engineering Journal*, **61**(10), 7953–7978.

[18] František Bartoš. (2022). *RoBSA: An R Package for Robust Bayesian Survival Analyses.* R package version 1.0.1, 2022.