Survival Analysis Method with R using Breast Cancer dataset

Wenhui Wang B00843110

Instructed by Dr.Lam Ho

Last compiled on December 17, 2022

List of Figures

1 Age Distribution of breast cancer patients 5
2 Kaplan-Meier survival rate curve 6
3 Kaplan-Meier survival curve divided by age 60. 7
4 Random Forest survival rate curve 10
5 The Nelson–Aalen estimator survival curve 11
6 Cox Proportional Hazards model survival curve. 12
7 Martingale residual test plot. 12
8 Martingale deviance residual test plot. 13
9 Schoenfeld test plot of age 14
10 Schoenfeld test plot of subtype 14
11 Cox Proportional Hazards model with multivariant. 15
12 Normal Q-Q plot of log-normal 17
13 Compare of kaplain, cox and random forest model 18

Contents

1 Introduction **4**

2 Data **4**

3 Methods **4**
3.1 Nonparametric Model 4
3.1.1 Kaplan-Meier Analysis 5
3.1.2 Random Forest Model 9
3.1.3 The Nelson–Aalen estimator 10
3.2 Cox Proportional Hazards Model 10
3.2.1 Martingale Residual Test 11
3.2.2 Schoenfeld Test 13
3.3 Parametric Models 15
3.3.1 Accelerated Failure Time Model (AFT) 16

4 Discussion **17**
4.1 Compare of the most popular three model-kaplain, cox and random forest model 17
4.2 Compare between AFT and PH model 17
4.3 Compare parametric and nonparametric 19

5 Reference **20**

1 Introduction

Survival analysis is a statistical method utilized to determine the time elapsed between two events. Accurately, the two events are a well-defined origin time and end time of a particular occurrence. It is also known as "Time to Event" Analysis and can be applied in many fields, including medicine, engineering, economics, etc.

Time-to-event data can be found in our life, for example, breast cancer survival rate after surgery. We need to think about the observed object's time and outcome for these events simultaneously [16]. Because time-to-event data includes censoring, a critical distinction between this type of data and cross-sectional data, conventional statistical approaches cannot be easily applied—that is why we need survival analysis. Censoring refers to incomplete data. It occurs when only a portion of the study subjects' exact event times are known, and the remainder is only known to exceed a specific value. The basic types of censoring include right, left, and random. In survival analysis, individuals are frequently followed not from time zero (in the relevant time scale, typically age), but only from a later entry time, as in epidemiological applications (conditional on survival until this entry time) [10]. As a result, the survival data are subject to left truncation and right censoring. We usually conduct univariate, bivariate, and multivariate analyses for quantitative research to test the research hypothesis; however, those methods are invalid in time-to-event data analysis due to censoring. Instead, researchers should analyze such data using a statistical model known as survival analysis [5]. The dependent variable in any survival analysis typically contains two pieces of information, distinguishing it from conventional statistical dependent variables. The first is a continuous variable that records the time (days, months, quarters, or years) the study subjects undergo a change process. The second is a dichotomous variable that indicates a change of state (i.e., going from state 1, "alive," to state 0, "dead"). This paper will introduce the principles and applications of several survival methods in a way that most people can understand [11].

2 Data

The data for this report was downloaded from the cBio Cancer Genomics Portal, a comprehensive, accessible resource for interactive exploration of multilevel cancer genomics data sets. In this paper, we will use Breast Invasive Carcinoma TCGA PanCancer data set as an example to illustrate survival analysis methods. This data includes 1084 patients' information, including sex, overall survival status, cancer type, ethnicity, race, tumour type, subtype and so on. In this data set, 1072 people are female, 12 are male, and 933 are still living. I choose this data set because it includes two important pieces of information on survival analysis: continuous variable as survival days and dichotomous variable as alive or dead. So it is proper to use the survival method to analyze. We can find the age distribution of the breast cancer data in Figure 1.

All data is manipulated and assessed in R, and the functions used to do so are derived from existing packages: survival, ranger, ggplot2, vctrs, dplyr, survmier, ggfortify, data.table, vcd, DescTools, mice, tidyverse.

3 Methods

We can divide the survival method into three sections: a nonparametric model, a semi-parametric model, and a parametric model. Nonparametric model is that the data can be collected from a sample that does not follow a specific distribution [11].

3.1 Nonparametric Model

Nonparametric approaches do not make assumptions about the shape or form of the underlying population's parameters. Ordinal or nominal data means Variables with no quantitative value are frequently subjected to nonparametric statistics. They are used in survival analysis to depict the data by assessing the survival function, S(t), descriptive statistics, statistical models, inference, and statistical tests. We always use Nonparametric methods as the first step in survival analysis [5].

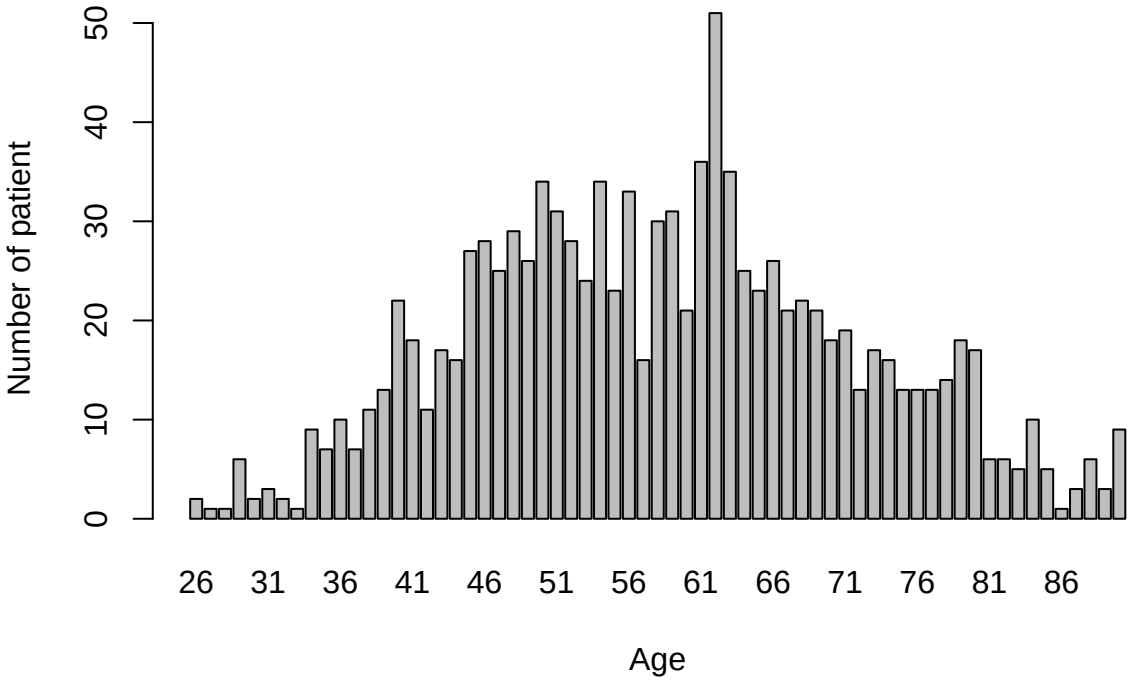

Figure 1: Age Distribution of breast cancer patients

3.1.1 Kaplan-Meier Analysis

Kaplan-Meier is a non-parametric estimator of survival proposed in 1958 by Edward L. Kaplan and Paul Meier, who independently came up with the same idea. Afterwards, John Tukey, who is the journal editor, convinced them to join in one paper, which was Kaplan-Meier's birth. Also, it is the most commonly used survival analysis.

Before we analyze Kaplan-Meier, each dataset will have three components: firstly, The study group, secondly, Their status at the end (alive or dead), and the last is survival time. Regardless of when they joined the research, the survival times for specific participants are ranked from the shortest to the longest. Using this technique, every subject in the group starts the analysis at the same point and continues to survive until one of them is dead. Survival time can be impacted by study participants who are uncooperative and refuse to stay in the study, or when some participants may not experience the event or die before the study is over, even though they would have if observation had continued, or when we lose contact with them in the middle of the study. These events are what we refer to as censored observations. Most survival data is right-censored, meaning the right side of observed survival time is incomplete. The actual survival time is longer than what we have [6].

The probability of surviving over a specified period while taking many tiny periods into account is known as the Kaplan-Meier survival curve. It is more like a declined step that does not rely on any parameter, not a smooth curve. That's why it is a non-parameter method. It entails calculating the likelihood of an event occurring at a specific time. We multiply these subsequent probabilities by any previously estimated probabilities to obtain the final analysis. That's why it is also called a "product limit estimate." The following formula can be used to calculate the likelihood of surviving at any given time:

$$\sum_{i:t_i<t} \frac{n_i - d_i}{n_i}$$

Which d_i denotes the number of deaths that happened at time t_i and n_i denotes the number of people who are known survivors at time ti [13].

Figure 2 represents the Kaplan-Meier survival curves. The horizontal axis (X-axis) shows time in days, and

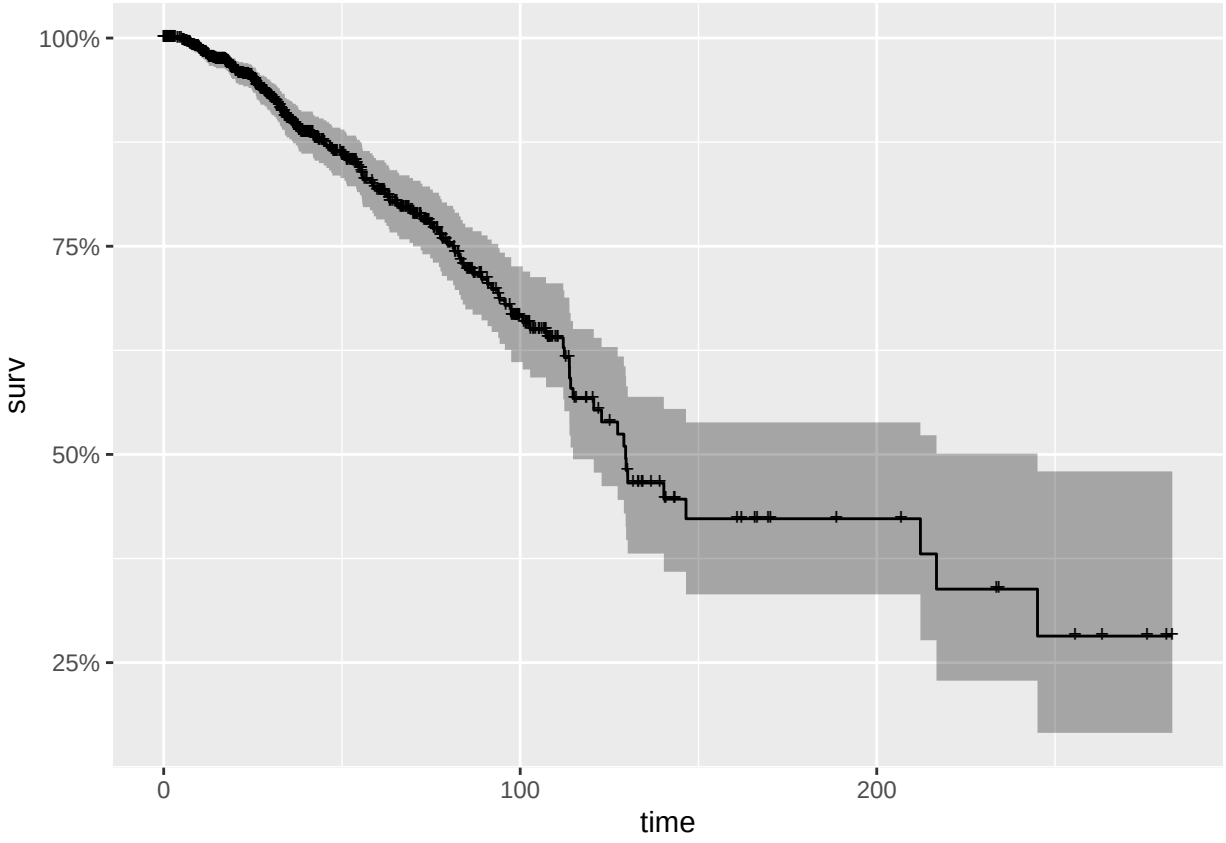

Figure 2: Kaplan-Meier survival rate curve

the vertical axis (Y-axis) shows the probability of survival. Every point on the curve represents the survival rate of patients. At zero, the likelihood of survival was 100 percent which means all of the participants were alive. With the increase of time, patients begin to die, and survival rates decline from 100 percent. A "+" in the graph indicates censoring data. Also, what we get is a confidence interval accompanying with survival curve. As time goes on, we find that the range of confidence intervals gets wider and wider. Because this is a cumulative result, the volatility and uncertainty of the first few periods will accumulate in the later period. In the last period, we find that the confidence interval is the largest.

We can divide the data by age to see the survival rate visually: creating a new data frame with a categorical variable AG with values LT60 and GT60, which describe patients below the age of 60 and above 60 We add SUBTYPE and PRIOR_DX into factor variables [14].

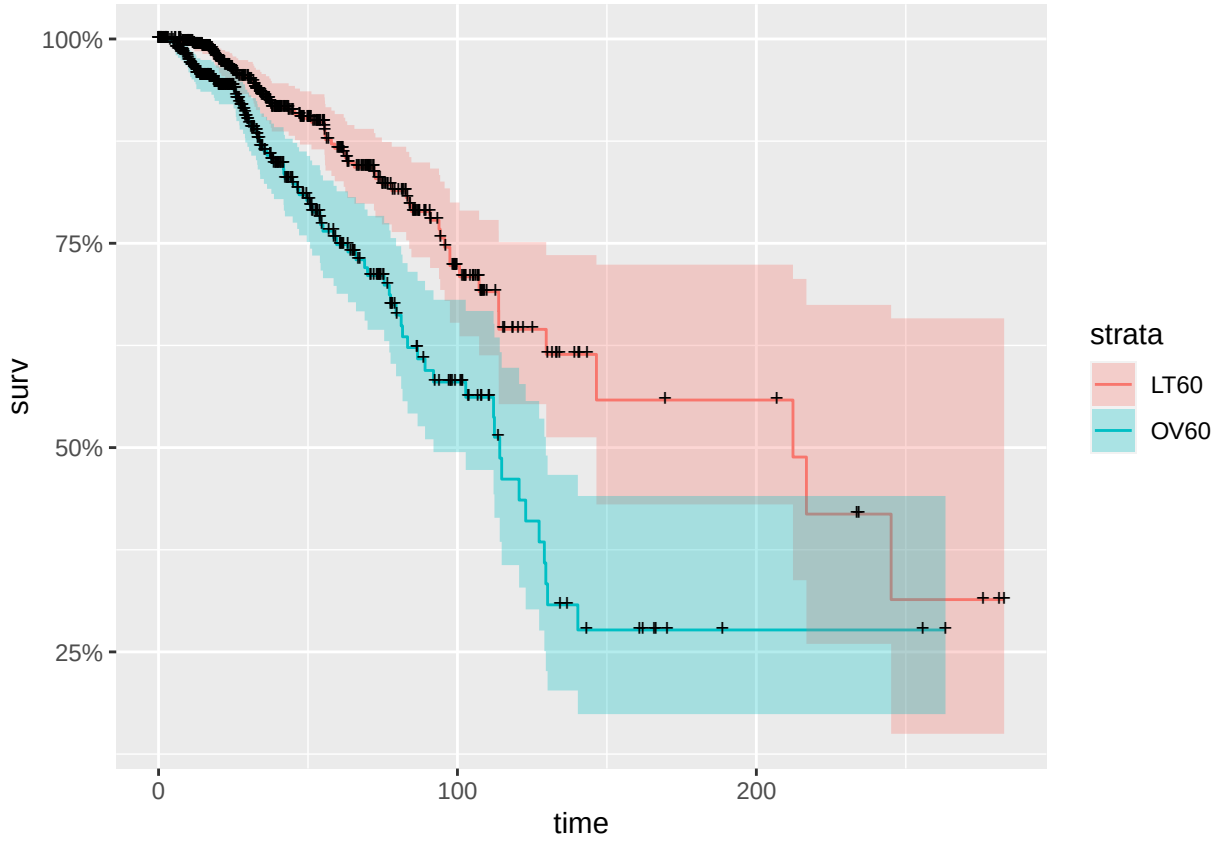

Figure 3: Kaplan-Meier survival curve divided by age 60.

```
##         records n.max n.start events    rmean se(rmean)   median   0.95LCL
## AG=LT60     578   578     578     66 181.9909  12.01651 212.2497 146.49702
## AG=OV60     506   506     506     85 137.3021  11.77351 114.1467  91.98803
##          0.95UCL
## AG=LT60       NA
## AG=OV60 129.5657
```

From the output, the median survival time was 114 days for people older than 60 years old and 212 days for those younger than 60 years old. We can conclude that the survival rate of patients more than 60 years old is lower than that of patients less than 60 years old at any time, so the group of patients more than 60 years old has a higher risk of death than the group of patients with less than 60 years old, and the survival rate is lower [14]. Younger people have an advantage when it comes to breast cancer. In Figure 3, the red curve represents the age lower than 60, and the blue curve represents the age over 60. The red curve is always on top of the blue curve, which also can indicate that younger people have a higher survival rate than older

people.

3.1.1.1 Long-rank Test

We always use the long-rank test and Wilcoxon test to determine whether survival distribution curves from two sub-populations differ significantly. The log-rank test is used to test the null hypothesis that there is no difference in the probability of an event between populations at any time. The test compares the entire survival experience of groups. It can be viewed as a test of whether or not the survival curves are identical (overlapping). The Kaplan-Meier method estimates survival curves for each group, which are then statistically compared using the log-rank test; however, the log-rank test is purely statistical. It cannot estimate the size of the difference between groups or a confidence interval. So, we must make some data assumptions [4].

First, let's make test hypotheses: Null hypotheses(H0): There is no difference in survival between the over 60-year-old and the under 60-year-old. Alternative hypotheses(HA): There is a difference in survival between the over-60-year-old and the under-60-year-old.

```
## Call:
## survdiff(formula = Surv(time_sur, status) ~ AG, data = patient_age)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## AG=LT60 578       66     91.1      6.93      17.6
## AG=OV60 506       85     59.9     10.54      17.6
##
##  Chisq= 17.6  on 1 degrees of freedom, p= 3e-05
```

From the result above, we can discover that the p-value is less than 0.05. So we reject the null hypothesis that there is a significant difference in survival between the two groups. We have sufficient evidence indicating that there is a difference in survival between the two groups.

3.1.1.2 Cochran-Mantel-Haenszel Test

Another method we compare with the long-rank test is the Cochran-Mantel-Haenszel Test. The difference between them is the method of estimating the variance. The Cochran-Mantel-Haenszel method generates an associative estimate between an exposure and an outcome after adjusting for or considering confounding. The method is applied to a binary outcome variable and a binary risk factor [11]. The data is stratified into two or more levels of the confounding factor. Essentially, we create two-by-two tables showing the relationship between the risk factor and the outcome at two or more levels of the confounding factor. Then we compute a weighted average of the risk ratios or odds ratios across the strata. We create two tables and divide them into binary outcome variables: 1. age above and under 60, 2. status is dead and living. We can produce eight values using the two conditions from breast cancer data [8].

```
##
##  Mantel-Haenszel chi-squared test with continuity correction
##
## data:  coch_test
## Mantel-Haenszel X-squared = 6.0841, df = 1, p-value = 0.01364
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.108053 2.214525
## sample estimates:
## common odds ratio
##          1.566464

##
```

```
##  Breslow-Day test on Homogeneity of Odds Ratios
##
## data:  coch_test
## X-squared = 8.1336, df = 1, p-value = 0.004345
```

By running the r code, we can discover that p-value is 0.01364 and X-squared is 8.1336, which is similar to the X-squared from the log-rank test.

3.1.2 Random Forest Model

Leo Breiman and Adele Cutler created the widely-used machine learning algorithm known as "random forest." It combines the output of various decision trees to produce a single outcome. They belong to ensemble learning and are classified as Bagging methods. Ensemble learning entails using many learners to improve the performance of any of them individually. These techniques combine a group of weak learners who only work well in their specific fields to create a more potent, aggregated model. What we discuss today are Random Forests that are made up of several individual Decision Trees.

We can divide "Random Forests" into two parts to help us understand it better: random and forest. Forest is an ensemble of many individual Decision Trees. Random is that every decision tree in the forest will sample a small part of the dataset for training. There are three training processes:

1. Creating a Bootstrapped Data Set for each tree. We must train a total of N decision trees, and select a random sample of the entire Data set for each tree. This is known as Sampling with Replacement or Bootstrapping: each data point is chosen randomly from the entire data set, and a data point can be chosen multiple times.
2. Create a decision tree with the corresponding data set, split on a random subset of variables or features at each node. Random Forest works best when the individual models (in our case, individual trees) are uncorrelated. This is accomplished in Random Forest by randomly selecting certain features to evaluate at each node. We avoid including features with highly high predictive power in every tree this way, resulting in a large number of un-correlated trees. This is the second round of randomness. We use random data as well as random features to build each tree. The more diverse the trees, the lower the variance and the better the model performs.
3. Repeat these three steps N times to make a massive forest with a wide variety of trees. This variety sets a Random Forest apart from a single decision tree.

Random Forest models combine the flexibility and power of ensemble models with the simplicity of Decision Trees. In a forest of trees, we forget about a tree's high variance and are less concerned with each element, allowing us to grow nicer, more giant trees with more predictive power than a pruned one. The algorithm will randomly take samples from the data set, and the randomness of the model is extreme, so it will not be overfitting. We can use a breast cancer data set to demonstrate random forests in R Studio: the ranger() function fits a Random Forests Ensemble model to the data [12].

From Figure 4, we can find the survival curve using the random forest method. We use the ranger() function to build a model for each variable and plot twenty random curves, and the black curves represent the global average for breast cancer patient data. Because it generates the curve randomly, the output is different every time.

```
##                                importance
## PERSON_NEOPLASM_CANCER_STATUS     0.1195
## AGE                               0.0292
## AJCC_PATHOLOGIC_TUMOR_STAGE       0.0150
## INFORMED_CONSENT_VERIFIED         0.0000
## SUBTYPE                          -0.0005
```

From the output above, we can find the importance of each variable. The ranger function label variable PERSON_NEOPLASM_CANCER_STATUS and AGE as the important variable, which is the same variable with the smallest p-value in the cox model that we will discuss later.

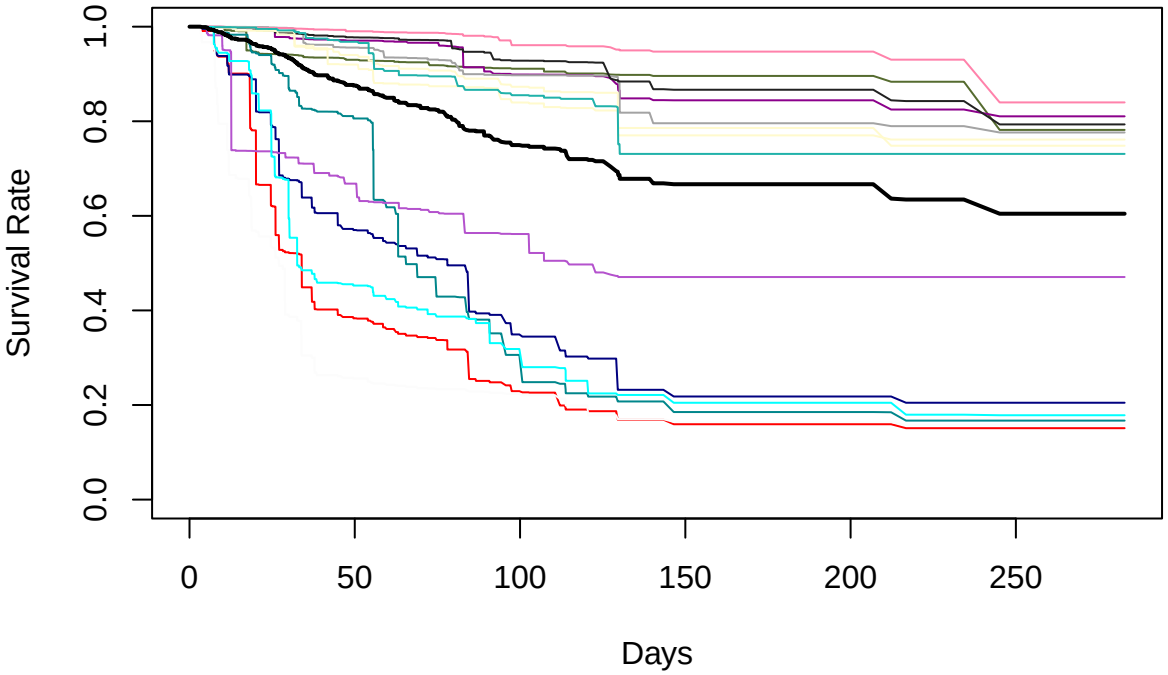

Figure 4: Random Forest survival rate curve

3.1.3 The Nelson–Aalen estimator

The Nelson-Aalen estimator can be used to estimate the cumulative hazard rate function. Because it does not need distributional assumptions, a critical application of the estimator is to check the fit of The Nelson–Aalen estimator for the cumulative hazard rate function is

$$\hat{A}(t) = \sum_{t_j=a} \frac{d_j}{r_j}$$

let d_j be the number of individuals who die at t_j and r_j is the number of individuals alive or not cencored [10].

We frequently want to estimate the survival distribution function S(t) = exp[A(t)], which represents the likelihood that an individual will be alive at time t. The Kaplan-Meier estimator can do this from right-censored and left-truncated survival data. The relationship A(t) = ln S(t) suggests that the cumulative hazard rate function can be estimated minus the logarithm of the Kaplan-Meier estimator [10]. Although this estimator will be numerically close to the Nelson-Aalen estimator, the latter is canonical from a theoretical standpoint. Furthermore, the Nelson-Aalen estimator can be used in various situations (see Figure 5), whereas the alternative estimator only applies to survival data.

In Figure 5, we can discern the cumulative hazard rate increase as time progresses.

3.2 Cox Proportional Hazards Model

Cox Proportional Hazards is the most commonly used semi-parametric regression model proposed by British statistician David. Cox in 1972. With survival events and survival time as dependent variables, this model can simultaneously analyze the influence of many factors on survival time.

Cox regression is crucial and widely applied for several reasons: The primary purpose of survival analysis is to study the relationship between variable X and survival function. The traditional method considers the regression equation when the survival function is affected by many factors. However, the data in survival analysis contains censored data, and the time variable t usually does not meet the requirements of normal

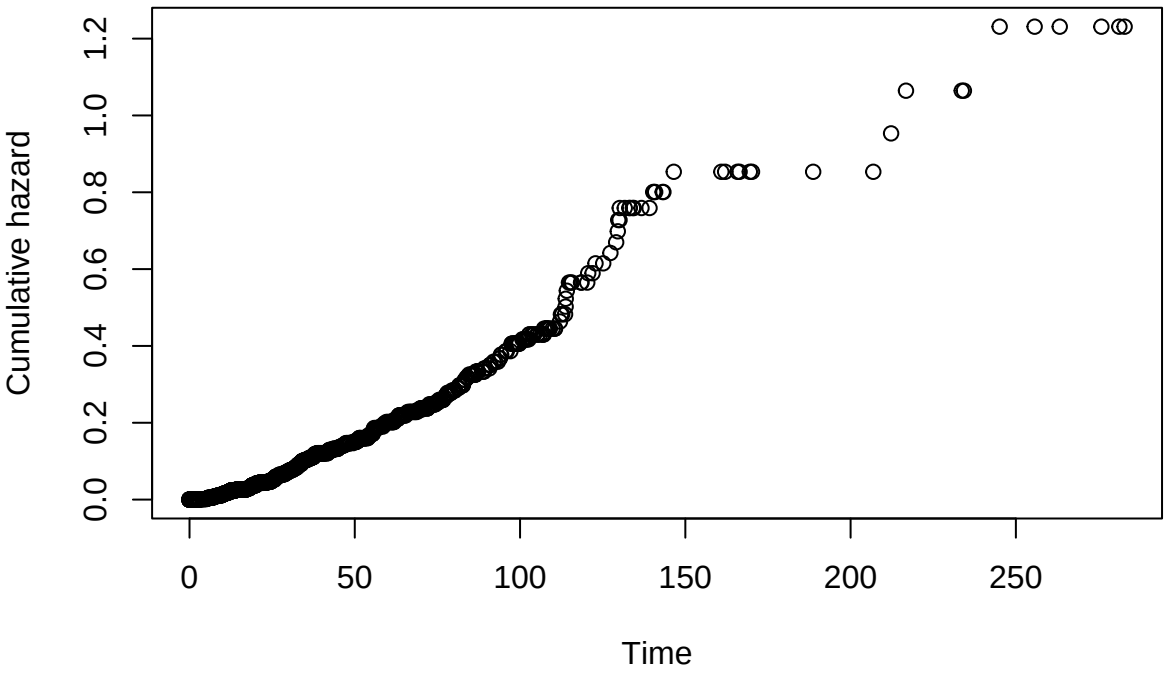

Figure 5: The Nelson–Aalen estimator survival curve

distribution and homogeneity of variance. So we introduce the cox proportional hazards model. An essential part of this method is to explore the risk factors that affect survival by influencing the risk of death per time unit, which is the hazard rate. People with different characteristics have different hazard rate functions at different times, usually expressed as the product of the underlying baseline hazard rate function and the corresponding covariate function. D.R.Cox put forward the Cox proportional hazards regression model, which is used as the dependent variable instead of directly investigating the relationship with X [10]. The basic form of the model is as follows: $h(t, X) = h_0(t) * f(x)$ Where $h(t, X)$ is the hazard rate function at time t and $h_0(t)$ is a baseline hazard function that does not have specific distributional assumptions, $f(x)$ is the parametric part that depends on each data set. That's why it is semi-parameter distribution. The most common use of $f(x)$ is : $hi(t, X) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)$

We will use age, subtype, race, and cancer status as four-factor covariates from the breast cancer data set to fit the cox regression model. We can see the cox model in Figure 6.

To evaluate the fitness of the Cox model, we can use Martingale and Schoenfeld test. Cox-specific goodness-of-fit tests, such as the Gronnesby and Borgan test and the Hosmer and Lemeshow prognostic index [15]. We are going to use the Schoenfeld Residuals Test, which is used to test the independence of residuals from time and, thus, the proportional Hazard assumption in the Cox Model. This test is equivalent to determining whether or not the slope of scaled residuals on time is zero. If the slope is not zero, the proportional hazard assumption is broken. In this test, each individual has a separate residual for each covariate. Schoenfeld residuals are the covariate value for individuals who failed minus the expected value. The PH assumption has been violated if the plot of Schoenfeld residuals against time shows a non-random pattern. To further test the independence of residuals and time, the residuals can be regressed against time.

3.2.1 Martingale Residual Test

Firstly, we can use Martingale residual test: we can check the assumption that every relation between x variable and log hazard is linear using the same method we use as linear regression. In Figure 7, I add a red line at residual equal to 0 and fit a smoother through these points.

Also, in Figure 8, we can check linearity using deviance residual which is predicted value versus deviance

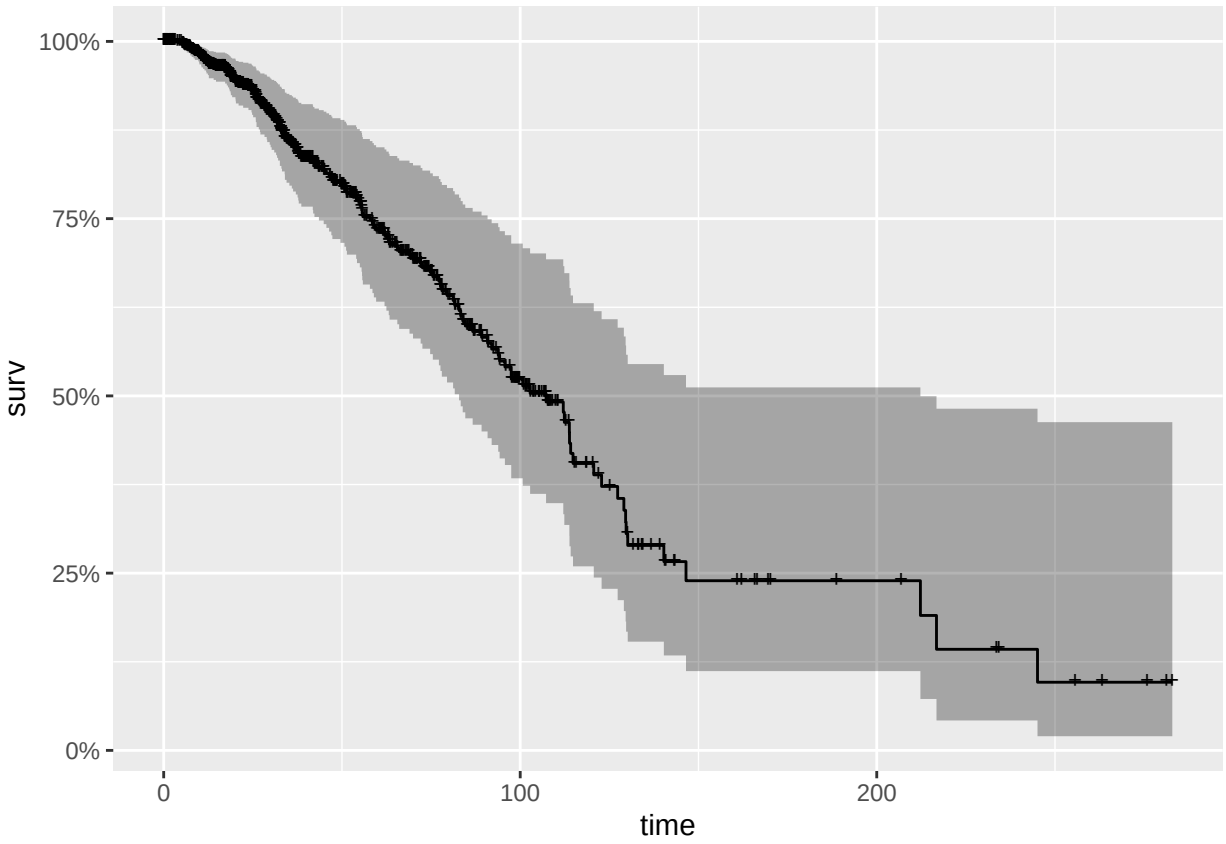

Figure 6: Cox Proportional Hazards model survival curve.

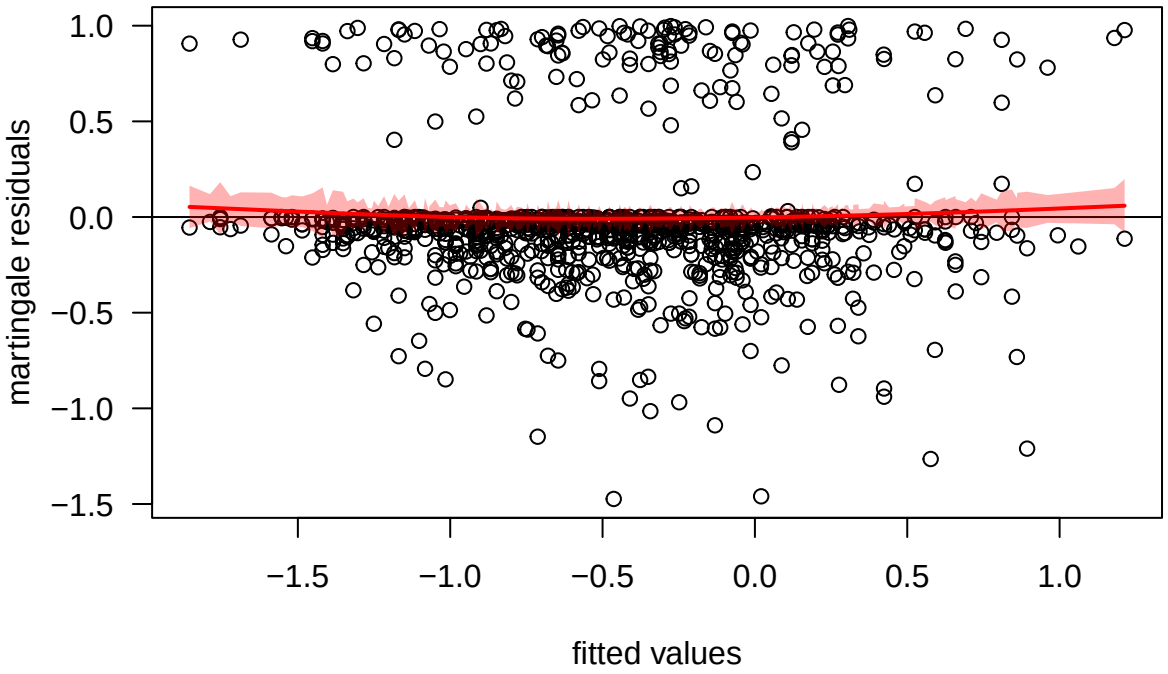

Figure 7: Martingale residual test plot.

residuals.

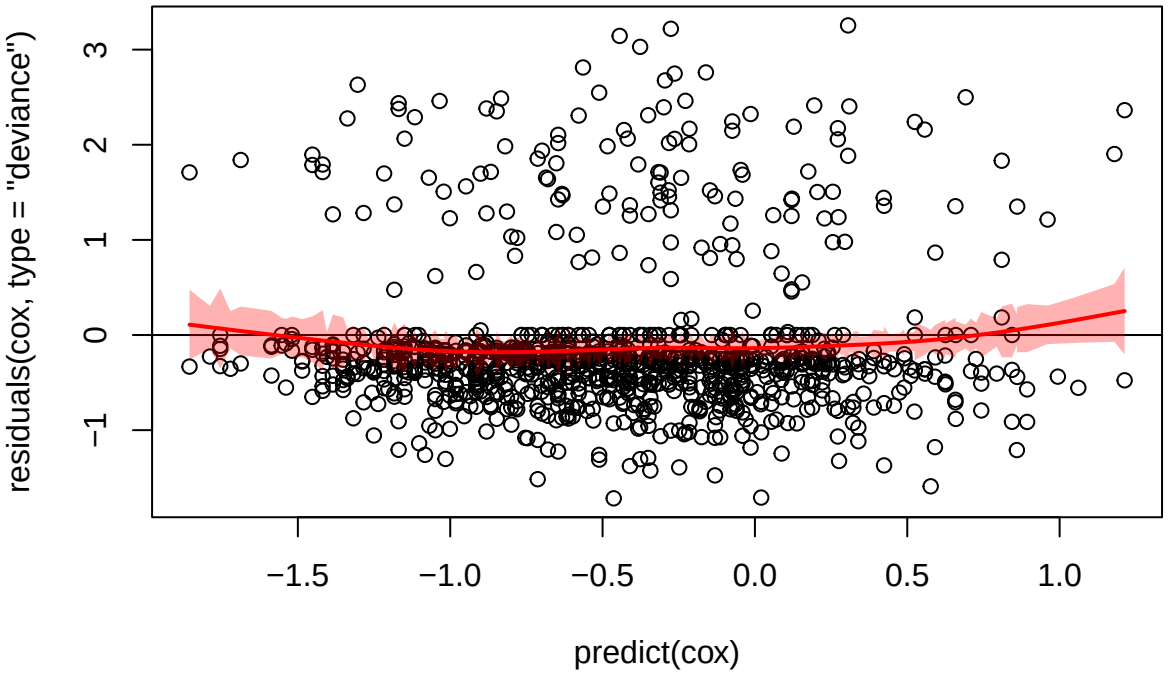

Figure 8: Martingale deviance residual test plot.

3.2.2 Schoenfeld Test

Then, we will check the proportional hazard assumption. We can use the Schoenfeld test, and the null hypothesis is that the cox proportional hazard model is proportional versus the alternative hypothesis that the cox proportional hazard is not properly. In survival package, zph() offers a practical way to test the proportional hazards hypothesis for each covariate included in a Cox regression model fit. The function cox.zph() correlates the corresponding set of scaled Schoenfeld residuals with time for each covariate to test for residual-time independence. Furthermore, it runs a global test on the entire model [1].

```
##          chisq df    p
## AGE     0.186  1 0.67
## SUBTYPE 6.290  5 0.28
## GLOBAL  7.465  6 0.28
```

From the output above, all p-values are larger than 5%. Thus, the test is not statistically significant for each covariate, and the global test is also not statistically significant. Therefore, we fail to reject the null hypothesis that we can assume proportional hazards.

The function ggcoxzph() generates graphs of the scaled Schoenfeld residuals against the transformed time for each covariate, allowing us to visualise it clearly [1].

In the Figure 8 and Figure 9, the solid line in the middle, which fits the smoother through points, means that if we allowed the hazard ratio to change over time, how much would the change produce? And the y-axis shows how much change it will be. It is zero, which means no change. The dashed line is the confidence interval around that. We can find that the solid line is a smoothing spline fit to the plot, with the dashed lines representing a +/- 2-standard-error band around the fit. We can add a red line at zero to see how often the zero is contained. And we can see that it is sometimes on zero about one-third time and most of the time above zero [14].

Also, we can conduct the cox model by multivariates. Figure 11 shows how the effects of the covariates

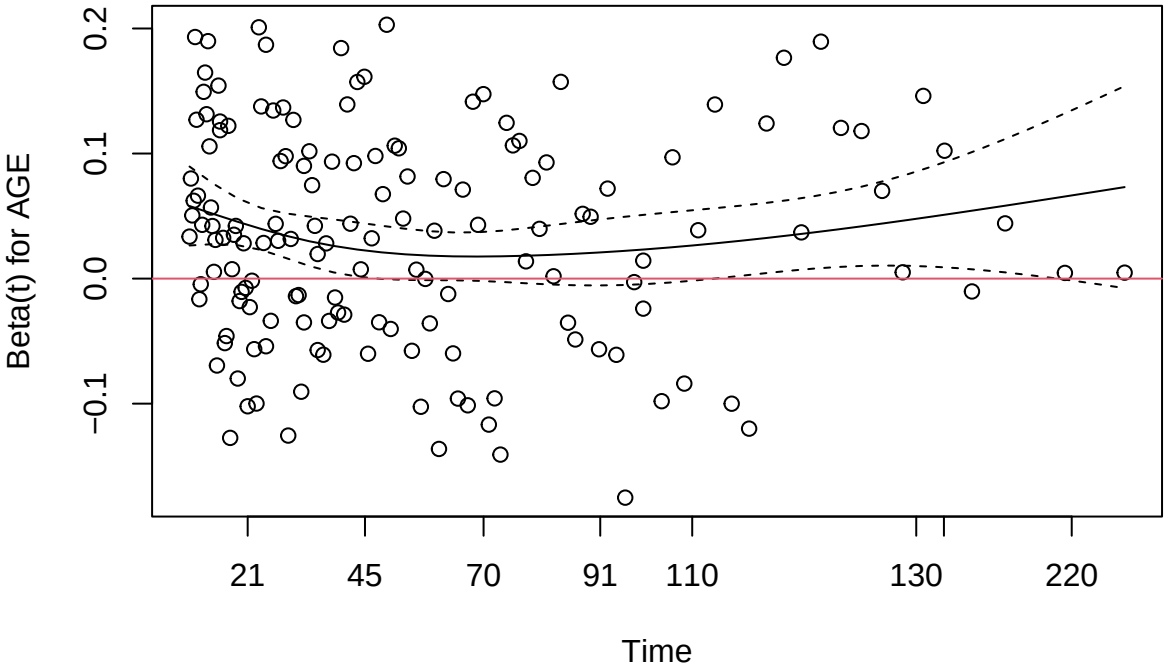

Figure 9: Schoenfeld test plot of age

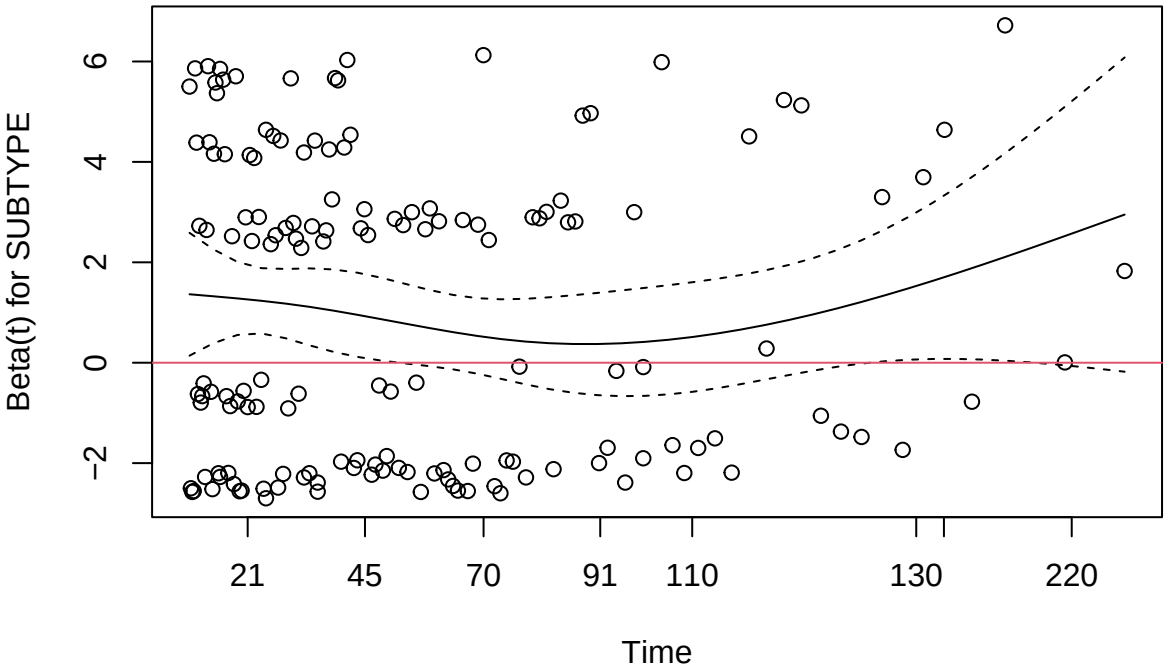

Figure 10: Schoenfeld test plot of subtype

change over time. What's more, we can see the p-value of each covariate in the output using the summary function [14].

```
##                   slope      coef se(coef)      z        p
## Intercept     -0.012900 -3.05e-03 1.33e-03 -2.290 2.18e-02
## AGE            0.000427  9.42e-05 1.75e-05  5.380 7.28e-08
## S_Tumor Free  -0.003140 -1.25e-03 5.85e-04 -2.130 3.30e-02
## S_With Tumor   0.029100  8.14e-03 1.45e-03  5.600 2.18e-08
## T_BRCA_Basal   0.002630  2.89e-04 8.93e-04  0.324 7.46e-01
## T_BRCA_Her2    0.006160  8.24e-04 1.15e-03  0.716 4.74e-01
## T_BRCA_LumA   -0.005670 -1.21e-03 7.94e-04 -1.520 1.27e-01
## T_BRCA_LumB   -0.001920 -4.05e-04 8.83e-04 -0.459 6.46e-01
## T_BRCA_Normal  0.002070  1.62e-04 1.25e-03  0.130 8.97e-01
##
## Chisq=71.77 on 8 df, p=2.18e-12; test weights=aalen
```

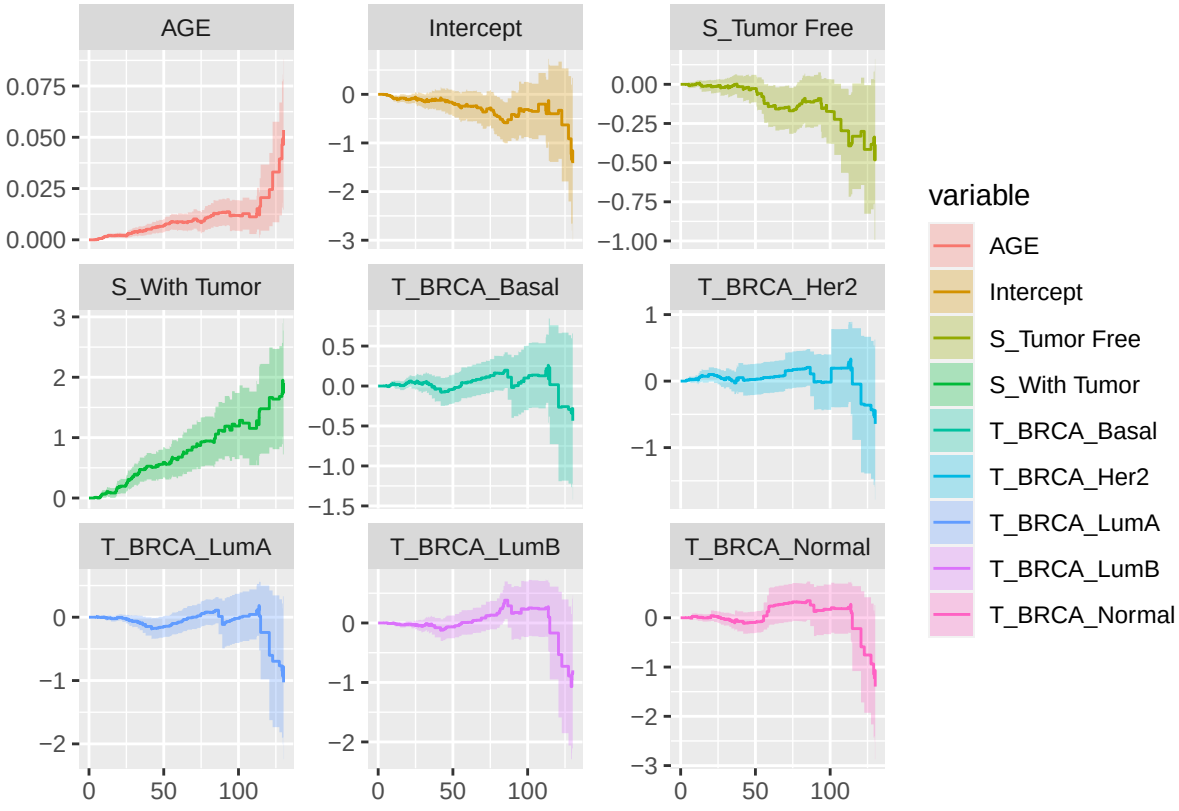

Figure 11: Cox Proportional Hazards model with multivariant.

3.3 Parametric Models

The modern survival analysis derives from the parametric models, an extension of the traditional ordinary least squares (OLS) regression. Understanding the main characteristics of parametric models can help comprehend the fundamental concepts of survival analysis and essential statistical concepts in general. They can run models with left-hand and interval censorings, and when the assumptions about the survival distribution are tenable, the model's estimates are usually reasonable, unbiased, and efficient.

Parametric distributions are often referred to as non-normal parametric. In survival analysis, parametric distributions such as exponential, Weibull, Gompertz, standard gamma, generalized gamma, lognormal, and log-logistic distributions are used. All of these models are in the form of an OLS regression. $\Upsilon = X\beta + e =$

$\beta_0 + \beta_1 X_1 + \beta_k X_k + e$Because we usually do not know which type of distribution is best suited to the data, the first task of running a parametric model is to determine its suitability of using a parametric model.

3.3.1 Accelerated Failure Time Model (AFT)

For time to event variable T, the Accelerated failure time model proposed that $Y_i = logT_i = X_i'\beta + W_i = \eta_i + W_i$we fit the mean with the covariate interest $X_i'\beta$ and Wi is the error term. According to the distribution we specify for W, we will obtain a different model, but all will have the same general structure, as described by the framework above [9]. For instance, assuming W ~ (0,σ^2)that Y follows a normal distribution is equivalent to assuming that T follows a log-normal distribution. However, most of the time, we will be focusing on Weibull distribution, which has extreme value and give us a more believable hazard function [16]. We will introduce some other parametric models and discuss their use of them later.

1. Exponential Model

The main characteristic of an exponential model is that the study time T has a constant hazard rate over the study window; that is, the rate of change of this distribution does not change over time, or h(t) = l for 0 £ t 1, where l is a constant [2].

$logT_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \sigma\epsilon_i$ The main benefit of this model is that it is both a proportional hazards model and an accelerated failure time model so that effect estimates can be interpreted as hazard ratios or time ratios [15]. The main disadvantage of this model is that assuming a constant hazard over time is frequently implausible [16].

2. Weibull Model

Weibull model has a hazard function of $h(t) = \lambda\gamma t^{-1}$, for $0<t<inf This function is determined by two parameters, lambda, and gamma. When gamma = 1, the hazard function is constant, implying that the survival times are distributed exponentially. As a result, an exponential distribution is a subset of a Weibull distribution. The survival function is e^{-x^γ}

3. Log-Logistic Model

The log-logistic distribution is an alternative model to the Weibull distribution. The log-logistic distribution has a fairly flexible functional form; it is one of the parametric survival time models in which the hazard rate can be decreasing, increasing, or hump-shaped, that is, increasing and then decreasing. Its survival function is $S(x) = (1 + ax^b)^{-1}$.

4. Log-normal Model

The log-normal distribution is a probability distribution derived from a continuous random variable and transformed from a normal distribution. When Y=ln(X) is normally distributed, then X is log-normally distributed, with meanμ and varianceσ.

We can use the AIC function to see which model is the most appropriate, and from the table, we can see that the Weibull distribution has the minimal value of AIC. We can say that it best fits the breast cancer data. Also, because the Exponential is nested in Weibull, we can use ANOVA to check whether Weibull or Exponential is more suitable. From the table, we can see the p-value is smaller than 0.05, which means Weibull is more appropriate.

```
##             df      AIC
## survregExp  13 1893.470
## survregWEI  14 1841.818
## survregLOGN 14 1847.029
## survregLOGL 14 1843.771
```

```
##                                                 Terms Resid. Df    -2*LL Test
## 1 AGE + SUBTYPE + RACE + PERSON_NEOPLASM_CANCER_STATUS      1071 1867.470
## 2 AGE + SUBTYPE + RACE + PERSON_NEOPLASM_CANCER_STATUS      1070 1813.818    =
##   Df Deviance     Pr(>Chi)
```

```
## 1 NA       NA           NA
## 2  1 53.65259 2.392676e-13
```

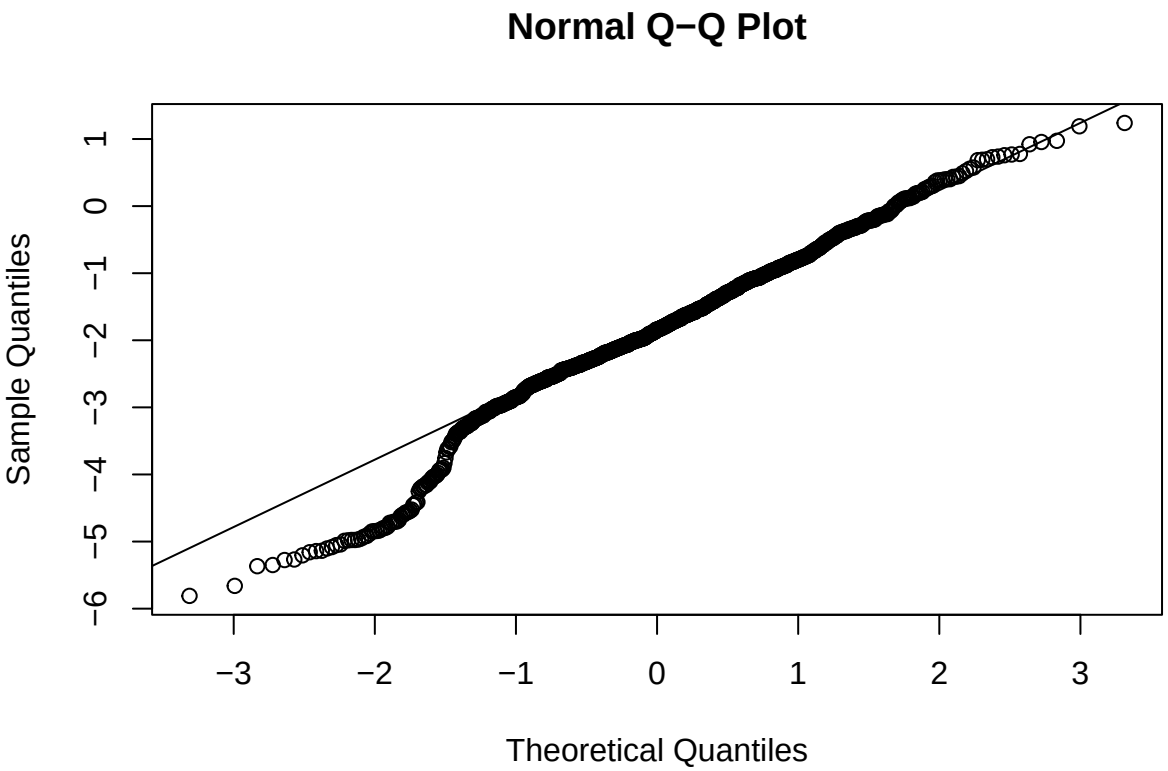

Figure 12: Normal Q-Q plot of log-normal

For log-normal distribution, we can use the Q-Q plot to check the normality shown in Figure 12. The bottom end of the Q-Q plot deviates from the straight line, but the upper is not. So we can say it is left-skewed, and we do not use log-normal for our dataset.

4 Discussion

4.1 Compare of the most popular three model-kaplain, cox and random forest model

In Figure 13, we can see the different survival lines of the three models. Kaplan-Meier model can provide an average overview of the model and does not have too many limitations. However, it cannot conduct multivariates data. We can use the Cox Proportional Hazards model, which can adjust for multiple risk factors simultaneously. The random forest model requires a lot of computational power as well as resources because it builds a lot of trees and then combines their outputs. We can use it as an alternative model when we perform survival analysis.

4.2 Compare between AFT and PH model

The Cox proportional-hazards model is the most commonly used survival analysis technique due to its fewer assumptions about the baseline hazard function. The formulation of the accelerated failure-time model allows the derivation of a time ratio, which is more interpretable than a ratio of two hazards. The AFT model does not require the PH assumption, which is rarely met in the Cox PH model. It also covers a broader range of survival time distributions and produces more powerful estimates than the Cox PH model. As a result, the AFT model is more appealing in various ways [3].

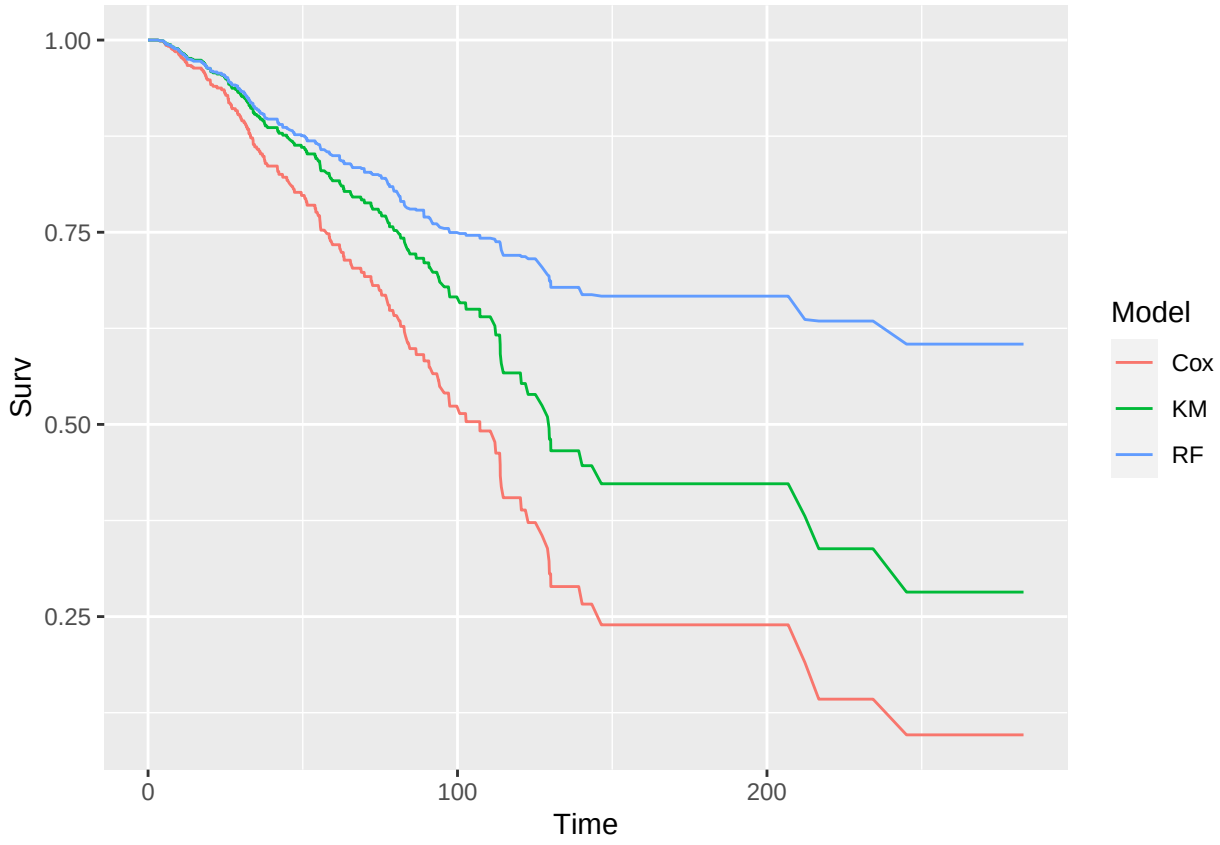

Figure 13: Compare of kaplain, cox and random forest model

An accelerated failure time model (AFT model) provides an alternative to proportional hazards models, which are commonly used. The effect of a covariate is the most significant difference between the AFT model and a proportional hazards model. However, the proportional model assumes that it is to multiply the hazard by some constant [9].

4.3 Compare parametric and nonparametric

Compared to the cox model and Kaplan model, The are some disadvantages of parametric models: they cannot be used to analyze time-varying covariates, and they require prior knowledge about the nature of the survival distribution being analyzed; if such information is not available, the user must assume that the empirical distribution being analyzed is the same distribution suggested by the parametric model; and when such an assumption is not valid, and the actual distribution is not the same kind of distribution as suggested by the parametric model, the user must assume that the empirical distribution being analyzed In practice, parametric models have been replaced by the Cox model for these reasons. The benefit criteria used in statistical analysis are not entirely clear. As a test of statistical conclusion validity, researchers frequently argue that the method must fit the research question and that assumptions must always be met. We frequently have options regarding statistical methods, and our choices should be tailored to the data situation [5].

5 Reference

[1] Cox proportional-hazards model. STHDA. (2022). Retrieved from http://www.sthda.com/english/wiki/cox-proportional-hazards-model

[2] Elfaki, F.(2017).Exponential Model for Survival Analysis. IJCSNS International Journal of Computer Science and Network Security. Retrieved from http://paper.ijcsns.org/07_book/201712/20171212.pdf

[3] Faruk, A. (2018). The comparison of proportional hazards and accelerated failure time models in analyzing the first birth interval survival data. Journal of Physics: Conference Series, 974, 012008. https://doi.org/10.1088/1742-6596/974/1/012008

[4] Finnstats. (2021). Log rank test in R-survival curve comparison: R-bloggers. Retrieved from https://www.r-bloggers.com/2021/08/log-rank-test-in-r-survival-curve-comparison/

[5] Guo. (2010). Survival analysis. Oxford University Press.

[6] Kishore, J., Goel, M. K., & Khanna, P. (2010). Understanding survival analysis: Kaplan-Meier estimate. International Journal of Ayurveda Research, 1(4), 274. https://doi.org/10.4103/0974-7788.76794

[7] Kleinbaum, & Klein, M. (2011). Survival Analysis: A Self-Learning Text, Third Edition. Springer New York.

[8] LaMorte, W. (2022). The Cochran-Mantel-Haenszel Method. Confounding and Effect Measure Modification. Retrieved from https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704-ep713_confounding-em/BS704-EP713_Confounding-EM7.html

[9] Mustefa, Y. A., & Chen, D.-G. (2021). Accelerated failure-time model with weighted least-squares estimation: application on survival of HIV positives. Arch Public Health. https://doi.org/10.1186/s13690-021-00617-0

[10] Nelson–Aalen estimator - Faculty of Medicine and Health Sciences. (2022). Retrieved from https://www.medicine.mcgill.ca/epidemiology/hanley/c609/material/NelsonAalenEstimator.pdf

[11] Parmar, & Machin, D. (1995). Survival analysis : a practical approach. J. Wiley.

[12] Pickett, K. L., Suresh, K., Campbell, K. R., Davis, S., & Juarez-Colunga, E. (2021). Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. BMC Medical Research Methodology, 21(1). https://doi.org/10.1186/s12874-021-01375-x

[13] Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C., Nussenbaum, B., & Wang, E. W. (2010). A practical guide to understanding Kaplan-Meier curves. Otolaryngology–head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery, 143(3), 331–336. https://doi.org/10.1016/j.otohns.2010.05.007

[14] Rickert, J. (2017). Survival Analysis with R. Retrieved from https://rviews.rstudio.com/2017/09/25/survival-analysis-with-r/

[15] Time-to-event data analysis. (2022). Retrieved from https://www.publichealth.columbia.edu/research/population-health-methods/time-event-data-analysis

[16] Wei. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. Statistics in Medicine, 11(14-15), 1871–1879. https://doi.org/10.1002/sim.4780111409