SURVIVAL ANALYSIS OF BREAST CANCER MORTALITY BASED ON THE SEER DATABASE

by

JIAYANG HE

Supervised By

Dr. Cindy Feng

Dr. Edward Susko

A thesis submitted to the
Department of Mathematics and Statistics
in conformity with the requirements for
Concentrated Honor in Statistics

Dalhousie University
Halifax, Nova Scotia
Jan, 2024

Abstract

Survival analysis is a statistical method used to analyze the time until an event of interest occurs, commonly applied in medical research, epidemiology, engineering, finance, and many other fields where the focus is on understanding the time until an event happens. The primary objective of this study is to conduct survival analysis aimed at identifying risk factors associated with the time from initiation of treatment to death among individuals diagnosed with breast cancer. Methods employed include Kaplan-Meier analysis, log-rank test, and the Cox proportional hazards model. Stepwise Akaike Information Criterion (AIC) is utilized for model selection to identify the final multivariable model. Schoenfeld residuals and Martingale residuals are employed to assess the validity of model assumptions. Additionally, Chi-square test and Wilcoxon rank sum test is utilized to remain focus on the modeling techniques related to survival analysis. Some potential issues, such as the limited sample size and violations of the proportional hazards assumption, are present in the final model. Plausible solutions to address these challenges are discussed.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Survival analysis, also known as "Time to Event" Analysis, is an analysis method that focuses on the time between the start of a particular event and a subsequent event [1]. One of the most common practices of survival analysis is to analyze the time from treatment to death but it is also widely utilized in other areas, such as economics, engineering, etc.

The main purpose of this paper is to conduct a survival analysis to study the duration between initiation of treatments and death among individuals diagnosed with breast cancer based on SEER data. Factors such as race, age, marital status, tumor status, and hormone stage are considered in the analysis. There are several reasons for choosing survival analysis as the major analysis method. Firstly, the dataset includes censored data, where the outcome of interest is only partially known, due to various reasons, such as: a) A patient has not yet encountered the pertinent outcome, such as relapse or mortality, by the study's conclusion, b) A patient becomes unavailable for follow-up during the study duration, c) A patient undergoes an alternate event that precludes further follow-up [2]. Moreover, the survival time is most unlikely to be normally distributed [1]. Given the unique characteristics, survival analysis method is the most appropriate approach for the study described in this paper.

In addition to the basic data exploration and generations of descriptive statistics, two major statistical models and tests in survival analysis will be utilized to explore the SEER breast dataset. Firstly, the Kaplan-Meier Survival Curves will be used to estimate and visualize survival probabilities stratified by covariates of interest. Next, the Cox Proportional Hazards (PH) Model will be employed to assess the relationship between covariates and breast cancer mortality and interpret the model coefficients to understand the impact of each covariate on the hazard rate. Moreover, the step AIC technique and likelihood ratio tests will be utilized to select the most appropriate final cox regression model. Then, to examine model assumptions, martingale, and Schoenfeld residuals are used.

Chapter 2

Data

This dataset, obtained from the November 2017 update of the Surveillance, Epidemiology, and End Results (SEER) Program by the (National Cancer Institute) NCI of the United States, provides valuable insights into population-based cancer statistics. The study specifically focused on female patients diagnosed with infiltrating duct and lobular carcinoma breast cancer (SEER primary sites recode NOS histology codes 8522/3) between 2006 and 2010. The dataset contains basic information from 4024 subjects, which includes age, race, marital status, and cancer information such as T stage, N stage, etc. With survival month and status data available for subjects, the dataset is well suited for conducting survival analysis. Additionally, approximately 84.7% of the individuals in the dataset are censored , i.e., did not experience the death during the study period or were lost to follow-up before the death occurred. The dataset is from Zenodo, a multi-disciplinary open repository maintained by CERN [3]. R studio is the main programming language used to generate all the necessary results and data visualizations. Libraries used include survival, ggplot, ggfortify, etc.

Data Description:

- Age (Continuous): The age of the patient at diagnosis for this cancer. The code represents the patient's actual age in years.

- Race (Categorical): The race recode is determined by the race variables and the American Indian/Native American IHS link variable. This recode is designed to connect with populations categorized as white, black, and other, regardless of Hispanic ethnicity. (Levels: White, Black, Other (American Indian/AK Native, Asian/Pacific Islander)

- Marital Status (Categorical): The patient's marital status at the time of diagnosis for the reportable tumor. (Levels: Single (never married), Married (including common law), Separated, Divorced, Widowed)

- T Stage (Categorical): Based on the size of the tumor and the extent to which it has grown into neighboring breast tissue. The higher the T number, the larger the tumor and/or the more it may have grown into the breast tissue. (Levels: T1, T2, T3, T4)

- N Stage (Categorical): Based on the number of lymph nodes involved and how much cancer is found in them. The higher the N number, the greater the extent of the lymph node involvement. (Levels: N1, N2, N3)

- 6th Stage (Categorical):Describes invasive breast cancer (cancer cells are breaking through to or invading normal surrounding breast tissue) (Levels: IIA, IIB, IIIA, IIIB, IIIC)

- Grade (Categorical): A measurement of how much the cancer cells look like normal cells. (Levels:Grade I, Grade II, Grade III, Grade IV)

- A Stage (Categorical): The extent of cancer spread beyond the primary tumor site. (Levels:Regional, Distant)
- Tumor Size (Continuous): The patient's current tumor size.
- Estrogen Status (Categorical): Whether the cancer cells may receive signals from estrogen that tell them to grow. (Levels:Positive, Negative)
- Progesterone Status (Categorical): Whether the cancer cells may receive signals from progesterone that tell them to grow. (Levels:Positive, Negative)
- Regional Node Examined (Continuous): Records the total number of regional lymph nodes that were removed and examined by the pathologist.
- Regional Node Positive (Continuous): Records the exact number of regional lymph nodes examined by the pathologist that were found to contain metastases.
- Survival Month (Continuous): The survival months of patients from the start to the end of the study.
- Status (Categorical): Any patient that dies after the follow-up cut-off date is recoded to alive as of the cut-off date. (Levels: Alive, Dead)

Chapter 3

Methods

Time-to-event variable is the variable of interest in survival analysis. A distinct feature of time-to-event variable is that it is often incomplete. In survival analysis, when the duration of observation for a subject is shorter than the time until the event of interest occurs, it is termed 'right censoring.' This incompleteness arises from the event not being observed within the study's time frame. Due to the difficulty of estimating the mean and variance for censoring data, different approaches are developed: Non-parametric and semi-parametric methods [4]. Specifically, the non-parametric method does not require distributions to analyze specific distributions[5], while semi-parametric statistical methods combine both non-parametric components and parametric components[6].

3.1 Non-parametric Methods

Non-parametric statistical methods do not make many assumptions, or none at all, about the shape or characteristics of the population distribution from which the sample was taken.

3.1.1 Censoring and Survival Analysis Fundamentals

Censoring

Censored data refers to the condition in statistical analysis and data collection where the value of a measurement or observation is only partially known [7]. This typically occurred when certain aspects of the data are obscured, leading to various types of censoring:

Right Censoring: The most common type, where the observation ends before the event occurs. For example, if a study ends and a participant has not yet experienced the event being studied (like relapse or equipment failure), their data is right-censored. All censored data in this research is right-censored [8].

Left Censoring: When the event of interest has already occurred before the start of the observation period. For instance, if a study starts observing patients after they have already contracted a disease [8].

Interval Censoring: Occurs when the event happens within a certain time interval, but the exact time is unknown. This can happen in medical studies where patients are checked at intervals, and the exact time of change (like tumor growth) between checks is unknown [8].

Random Censoring: This assumes that the censoring of an observation is independent of the event time. It's a more general assumption and can be unrealistic in certain studies [8].

Hazard Function

The hazard function, also known as the hazard rate, is a fundamental concept in survival analysis . It describes the instantaneous rate of occurrence of the event of interest (like failure or death) at a given time, assuming that the event has not occurred up to that time. The hazard function is used to provide insights into the dynamics of event occurrence over time and can vary significantly depending on the type and conditions of the data being analyzed [9].

The formula of hazard function is given by:

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

where the variables are defined as follows:

$h(t)$ - Hazard function at time t, representing the instantaneous event rate.

Δt - A very small time interval.

T - A random variable representing the time until the occurrence of the event.

$\Pr(t \leq T < t+\Delta t \mid T \geq t)$ - Probability that the event occurs within the interval $[t, t + \Delta t)$, given that it has not occurred by time t.

A continuous failure rate is contingent on the presence of a failure distribution, denoted as $F(t)$. This function is a cumulative distribution function that outlines the likelihood of a failure occurring by or before time t, and is defined as follows

$$\Pr(T \leq t) = F(t) = 1 - S(t), \quad t \geq 0$$

Here, T symbolizes the time until failure. $S(t)$ denotes the survival function and it represents the probability that the time until an event (like failure, death, or another endpoint) is longer than a specific time t. The function $F(t)$ is computed as the integral of the failure density function $f(t)$, over the interval from zero to t:

$$F(t) = \int_0^t f(\tau)\, d\tau$$

Based on these definitions, the hazard function $h(t)$ is then defined by the ratio of the failure density function to the remaining probability of survival, which simplifies to:

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

This formulation provides insight into the rate at which failures occur as a function of time, taking into account the cumulative probability of failure and the probability of not having failed up to that point [10].

3.1.2 Kaplan-Meier Analysis

The Kaplan–Meier estimator is a non-parametric statistic used to estimate the survival function from lifetime data and it analyzes time-to-event data such as the time until a patient experiences a certain event, such as death [11].

The Kaplan–Meier analysis is a common method for survival analysis. One distinguished feature of survival analysis is censored observation. Censoring data is defined as a scenario within a study where only incomplete information is available for some subjects due to various reasons such as non-cooperation, dropouts, or the study's conclusion before the occurrence of the event of interest. These cases are known as right-censored observations. They offer partial insight since it's only known that these subjects have survived up until a certain last known point, and the precise timing of the event post-follow-up remains unknown. Although the exact event timing for these individuals is not determined, their data is still useful and should not be discarded, as it confirms survival up to the last observation point. These considerations are vital in survival analysis, as they contribute to the understanding of survival probabilities without skewing the overall findings. [12].

The Kaplan–Meier estimator is calculated as:

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

Where:

$\hat{S}(t)$ is the estimated survival probability at time t.

t_i are the distinct observed time points.

d_i is the number of events (e.g., deaths) observed at time t_i.

n_i is the number of individuals at risk of experiencing the event at time t_i, which means all subjects who have not yet experienced an event (like failure or death) at time t_i.

The estimator works by calculating the probability of survival at each distinct time point in the dataset. It does this by dividing the number of individuals who have not experienced the event of interest by the number of individuals at risk just before that time point. The probability of survival is then multiplied cumulatively across time points to estimate the overall survival function.

There are three assumptions for the Kaplan-Meier Analysis. Firstly, it is presumed that individuals whose data are censored at any given point maintain equivalent survival likelihoods to those who remain under observation. Secondly, the assumption is that the survival probabilities do not differ between participants enrolled at the beginning of the study and those who join later, which means that the survival times are supposed to be independent? Lastly, it is presumed that the events of interest occur precisely at the times recorded.

The Kaplan–Meier curve graphically represents the survival rate or survival function. Time is plotted on the x-axis and the survival rate is plotted on the y-axis. The steeper the curve, the higher event rate or death rate, which means a worse survival prognosis. Conversely, the flatter the curve, the lower event rate or death rate, therefore a better survival prognosis. The survival probabilities can be compared between groups if multiple curves are included in the graph. If curves are parallel to each other, the survival probabilities are similar between groups. If curves intersect or diverge, the survival probabilities are notably different [13].

3.1.3 Logrank Test

The logrank test is a non-parametric hypothesis test used for censored data in survival analysis to compare the survival distributions of two or more groups. The null hypothesis for the logrank test is that there is no distinction between the populations in terms of the likelihood of an event (in this case, a death) occurring at any given time, while the alternative hypothesis is there is distinction between the populations in terms of the likelihood of an event occurring at any given time [15].

To perform the logrank test, various assumptions are needed. Firstly, it assumes independent censoring, where the censored data is not related to the chances of the event of interest. Secondly, the test presumes that all groups being compared have the same survival distributions when the null hypothesis holds true. Thirdly, a consistent ratio of hazard rates between groups over time is presumed, an assumption known as proportional hazards. Additionally, the test assumes that event times are accurately recorded, and that there are no drop-outs or withdrawals other than those accounted for by censoring. Finally, the test assumes homogeneity in the event rates across the study period, with no fluctuations in the survival differences between the groups.

The logrank test statistic evaluates the hazard functions of two groups at each observed event time by comparing their estimated values. It is calculated by determining the observed and expected number of events in one of the groups at each event time and then combining these values to generate a comprehensive summary across all event times where an event occurs [14].

The logrank test statistic is defined as:

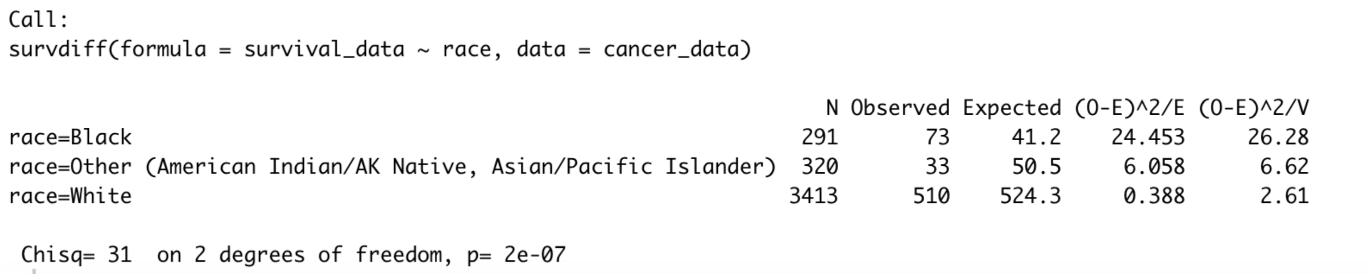
$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{3.1}$$

where:

$$
\begin{aligned}
X^2 &= \text{the chi-squared test statistic.}\\
O_i &= \text{the observed number of events in the group } i.\\
E_i &= \text{the expected number of events in the group } i,\\
k &= \text{the total number of distinct event times observed in the study.}
\end{aligned}
$$

Performing the logrank test is divided into five step. Firstly, divide the observation period into distinct time intervals or event times. Then, for each event time, calculate the observed number of events (e.g., deaths) in each group. Next, calculate the expected number of events in each group under the null hypothesis of no difference. This is typically based on the assumption that the hazard functions are the same for all groups. Moreover, compute the logrank test statistic, which measures the discrepancy between the observed and expected numbers of events across all event times. Lastly, assess the significance of the logrank test statistic using the chi-squared distribution with appropriate degrees of freedom [15].

As an illustration of logrank test, suppose we want to compare the survival probability among the ethnicity groups, i.e., white, black and other). Below shows the output using survdiff function in the survival package in R.



```
Call:
survdiff(formula = survival_data ~ race, data = cancer_data)

                                                                  N Observed Expected (O-E)^2/E (O-E)^2/V
race=Black                                                      291       73     41.2    24.453     26.28
race=Other (American Indian/AK Native, Asian/Pacific Islander)  320       33     50.5     6.058      6.62
race=White                                                     3413      510    524.3     0.388      2.61

 Chisq= 31  on 2 degrees of freedom, p= 2e-07
```

Here is the interpretation of key components in the output:

N: This column indicates the number of individuals in each racial group included in the analysis.

Observed: This column shows the actual number of events (deaths) observed in each racial group.

Expected: This column displays the expected number of events in each racial group under the null hypothesis of no difference between the groups. It's typically calculated based on the overall event rate in the entire sample.

$(O-E)^2/E$: This column presents the contribution of each racial group to the overall chi-squared statistic. It measures the discrepancy between the observed (O) and expected (E) numbers of events, standardized by the expected number of events.

$(O-E)^2/V$: This column is similar to the previous one but standardized by the total variance (V) of the observed minus expected counts across all groups.

Chisq: This value represents the overall chi-squared statistic, which quantifies the overall difference in survival distributions among the racial groups.

Degrees of freedom: This indicates the number of categories minus 1. In this case, there are three racial groups, so the degrees of freedom are 2.

p-value: This value indicates the statistical significance of the chi-squared statistic. A small p-value (e.g., 2e-07) suggests that there is strong evidence to reject the null hypothesis of no difference between the racial groups in terms of survival outcomes.

Additionally, the expected number of events for each group in the output can be calculated as:

$$\text{Expected number of events for group } i = \sum_j \left(\frac{n_{ij}}{n_j} \cdot d_j \right)$$

where the variables are defined as:

n_{ij} is the number of individuals at risk in group i just before time j.

n_j is the total number of individuals at risk in all groups just before time j.

d_j is the total number of events (such as deaths) observed at time j across all groups.

To calculate the expected number of events for a group across the entire study period, sum the products of the proportion of individuals at risk in the group and the number of events at each time point[16].

$$\text{Overall expected number of events for group } i = \sum_j \left(\frac{n_{ij}}{\sum_k n_{kj}} \cdot d_j \right)$$

This sum is taken over all time points j where events occur, providing the total expected number of events for group i under the null hypothesis that the survival functions are the same across all groups[16].

3.2 Cox Proportional Hazards Model

The Cox proportional hazards model is a statistical technique introduced by David Cox in 1972 that can be used for survival-time (time-to-event) outcomes on one or more predictors. Proportional hazards regression, often referred to as the Cox regression model, holds significance as a multivariable model, particularly in situations where the outcome involves the duration until a specific event occurs, such as infection or death. This model is commonly used to evaluate the relationship between the survival time of patients and one or more predictor variables.

A key advantage of using the Cox regression model in survival analysis is its ability to handle censoring data. Censored data occur when the event of interest has not occurred by the end of the study for some participants. These patients contribute data only up to the time they are censored, thereby influencing the overall estimation of the time to events within the population. It's important to recognize that in scenarios where there's no censorship, such as when all patients experience the event, a generalized linear regression model would be suitable for modeling the time to the event.

In Cox model, he response variable is the hazard function $h(t)$, which assesses the probability that the event of interest (in this case, death) occurred before t. The equation models this hazard as an exponential function exp of an arbitrary baseline hazard $h_0(t)$ when all covariates are all set at zero, and β is the regression coefficient of the covariate, x [17].

Here is the formula for the Cox proportional hazards model:

$$h(t \mid X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i)$$

where:

$h(t \mid X)$ is the hazard at time t for an individual given the covariate X.

$h_0(t)$ is the baseline hazard, which is the hazard function for an individual with the baseline level (typically zero) of the covariates.

$\exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i)$ is the exponential function of the linear predictor, which is the sum of the products of covariates X_i and their corresponding coefficients β_i.

β_i represents the coefficient for the i-th covariate, indicating the effect size of the covariate on the hazard.

X_i represents the value of the i-th covariate.

The $\exp(\beta_i)$ terms are referred to as hazard ratios, and they describe how the hazard changes in response to a one-unit change in the i-th covariate, assuming other covariates are held constant. The model's assumption of proportional hazards implies that the effect of the covariates on the hazard is multiplicative and does not change over time.

One of the model's core assumptions is the proportional hazards assumption, which indicates that the effects of the covariates on the hazard are multiplicative and remain constant over time. This implies that the hazard functions for different groups (strata) are proportional and do not cross over time.

Another key assumption of the Cox model is that the relationship between the natural logarithm of the hazard and each covariate is linear. This linear relationship can often be verified through residual plots, which help assess whether the model assumptions are reasonably satisfied in the given data.

Cox proportional hazards regression model

In the Cox Proportional Hazards (PH) model, the estimates for the parameters are derived by maximizing the partial likelihood . The partial likelihood, which is utilized for this purpose, is presented as follows:

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(\beta' X_i)}{\sum_{j \in R(t_i)} \exp(\beta' X_j)}$$

where:

$L(\beta)$ is the partial likelihood function.

β represents the vector of coefficients associated with the covariates.

$\beta' X_i$ is the dot product of the coefficients vector and the covariates vector for the i-th individual who experienced the event.

$R(t_i)$ denotes the risk set at time t_i, which is the set of all individuals who are at risk of the event just prior to time t_i. In other words, it represents all subjects who have not yet experienced an event (like failure or death) at time t_i.

The numerator $\exp(\beta' X_i)$ is the estimated hazard for the i-th individual who experienced the event.

The denominator $\sum_{j \in R(t_i)} \exp(\beta' X_j)$ is the sum of the estimated hazards for all individuals in the risk set at t_i.

The log partial likelihood function for the Cox proportional hazards model is expressed as the logarithm of the partial likelihood function, given by:

$$l(\beta) = \log L(\beta) = \sum_{i:\delta_i=1} \left[X_i\beta - \log \left(\sum_{j:Y_j \geq Y_i} \exp(X_j\beta) \right) \right]$$

where:

$l(\beta)$ is the log partial likelihood function.

$L(\beta)$ is the partial likelihood function.

β represents the vector of coefficients associated with the covariates.

X_i is the vector of covariates for the i-th individual.

Y_i is the observed survival time for the i-th individual.

δ_i is an indicator that is 1 if the i-th observation is uncensored and 0 otherwise.

The inner sum is taken over all individuals j whose survival time Y_j is greater than or equal to Y_i, signifying that they are at risk at time Y_i.

Cox and others have demonstrated that the partial log-likelihood from the Cox Proportional Hazards model functions can be considered as an ordinary log-likelihood. This similarity allows for the derivation of valid (partial) maximum likelihood estimates (MLEs) of the coefficients . As a result, it is feasible to estimate hazard ratios and confidence intervals using maximum likelihood methods that are traditionally applied to full likelihoods, though in this case, the calculations are based on the partial likelihood. The use of the partial likelihood is considered valid and reliable as long as there are no ties in the dataset, meaning no two subjects experience the event at precisely the same time. However, if ties do occur, calculating the true partial log-likelihood becomes considerably more complex and time-consuming, involving permutations of the data [18].

3.2.1 Model Selection

Model selection is a critical step in building Cox PH model to identify most relevant covariates associated with survival outcomes. Akaike information criterion [19] is a commonly used metric for selecting variables in survival analysis.

The Akaike Information Criterion (AIC) assesses the quality of statistical models by balancing the trade-off between goodness of fit and model complexity [20]. The AIC can be utilized in model selections by comparing the AIC values of different models fitted to the same dataset. Models with lower AIC values indicate a better balance between goodness of fit and model complexity. Researchers will select the model with the lowest AIC values as the most appropriate model for the data. In this case, the risks of overfitting and underfitting for the model will be eliminated, which maximizes the accuracy of the prediction model [21].

The AIC value is calculated using the following formula:

AIC = 2k - 2ln(L)

Where:

k = Number of parameters in the model

L = Likelihood of the observed data under the model

ln(L) = Natural logarithm of the likelihood function

The large parameter number(k) generally means a more complex model, which can fit the data better but might also lead to overfitting (i.e., fitting the noise in the data rather than just the underlying relationship).

The likelihood function measures how likely it is to observe the given data under a specific statistical model with certain parameters. A higher value of L indicates that the model with the chosen parameters is more likely to produce the observed data, which means that the model fits the data better with a high values of L. $ln(L)$ is used instead of L because the likelihood can be a very small number, making it difficult to work with computationally. The logarithm transforms the product of probabilities into a sum, simplifying the calculation [21].

In R, the step() function is commonly used for the context of model selection procedures, particularly in the context of the Cox proportional hazards model in survival analysis. The step() function performs backward step-wise regression by default, which is a method for iteratively removing variables from the model based on certain criteria, such as minimizing the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC).

In the research, the response variable is created by ’Surv()’ function in R. This function is used to create a survival object, which is a special type of data structure used in survival analysis to represent survival times and censoring indicators. The response variable is the survival object created by ’Survival Month’ and ’Status’ variables in the dataset, and it is named ’survival$_{d}ata'$.

The initial Cox proportional hazards model assumes all the covariates have an impact on the survival outcome. Then, in order to select the most appropriate model for the dataset, the step() function is utilized. The function with the lower AIC value is selected as the final model for the dataset.

3.2.2 Schoenfeld individual test

Schoenfeld residuals are a diagnostic tool used in survival analysis to assess the proportional hazards assumption in Cox proportional hazards models. Schoenfeld residuals are obtained by maximizing the partial likelihood function with respect to regression coefficient of the Cox model, which are defined as the difference between the observed and expected values of a covariate at each event time. If the assumption of PH holds, a plot of these residuals against ordered death times should look like a random walk without any pattern. [22].

The null hypothesis for the Schoenfeld individual test is that there is no violation of the proportional hazards assumption. In other words, the hazard functions of the groups being compared do not vary over time. The alternative hypothesis is that there is a violation of the proportional hazards assumption. The hazard functions of the groups being compared (e.g., treatment vs. control) vary over time, indicating that the proportional hazards assumption is not met.

3.2.3 Martingale Residual

Martingale residuals are used to assess the goodness-of-fit for a Cox proportional hazards model. Since the Cox model is a hazard model, traditional residuals don't apply. Martingale residuals are instead calculated as the difference between the observed number of events and the expected number of events, as predicted by the model, up to a certain time [23].

The formula for Martingale residuals is:

$$M_i = \delta_i - \int_0^{t_i} h_0(u) e^{\beta' X_i} \, du$$

where:

M_i is the Martingale residual for the i-th individual.

δ_i is the event indicator, which is 1 if the event (such as failure or death) has occurred for the i-th individual, and 0 otherwise.

t_i is the observed time until the event or censoring for the i-th individual.

$h_0(u)$ is the baseline hazard function at time u, which is unspecified in the Cox model.

β is the vector of coefficients estimated from the Cox model.

X_i is the vector of covariates for the i-th individual.

While martingale residuals are useful for assessing the overall fit of a model, their asymmetric nature can make it challenging to utilize them for identifying outliers.

Chapter 4

Results

The study focuses on the SEER Breast Cancer dataset encompassing 4,024 individuals, incorporating essential demographic and clinical particulars such as age, race, marital status, cancer stages, and hormone receptor status. Given the presence of censored observations—wherein 84.7% of the data does not have the event of interest occurring within the study period, survival analysis is an effective method to handle such incomplete data effectively.

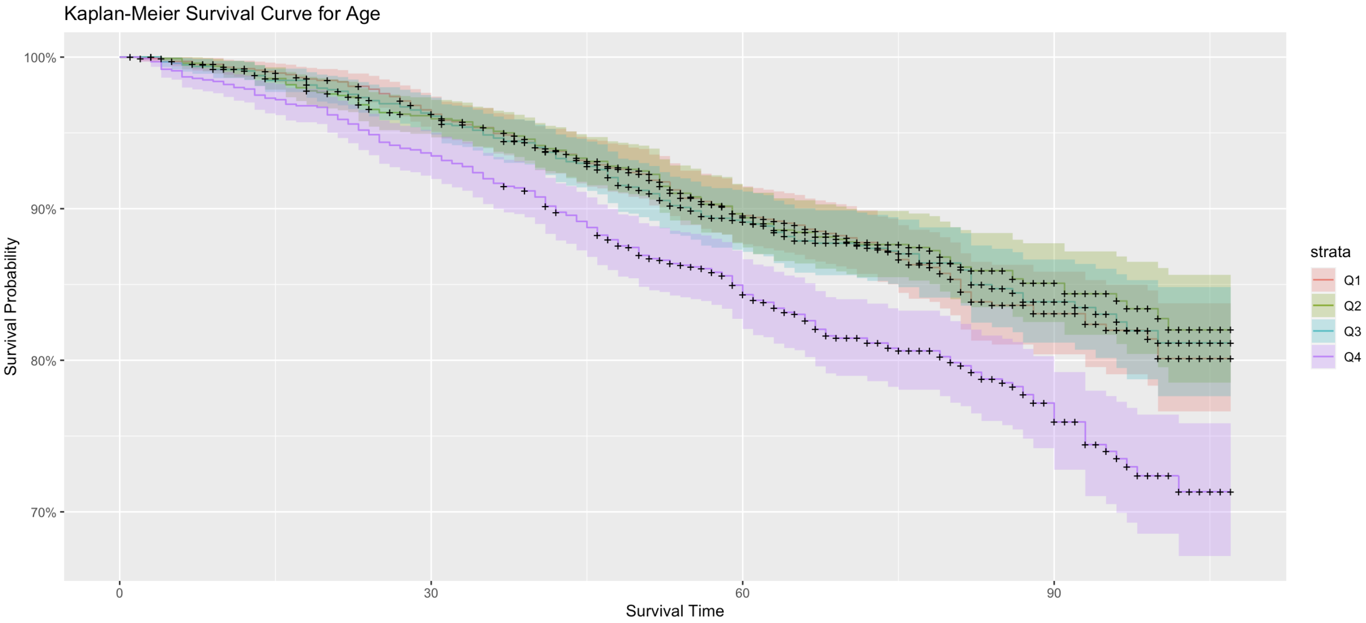
The primary goal of this research is to identify key determinants that significantly impact the survival probabilities of patients. To achieve this, a blend of non-parametric techniques, specifically the Kaplan-Meier Analysis and logrank test, alongside a semi-parametric approach, the Cox proportional hazards model, are employed.

This section is dedicated to presenting and interpreting the outcomes derived from these statistical methodologies.

4.1 Kaplan-Meier Analysis

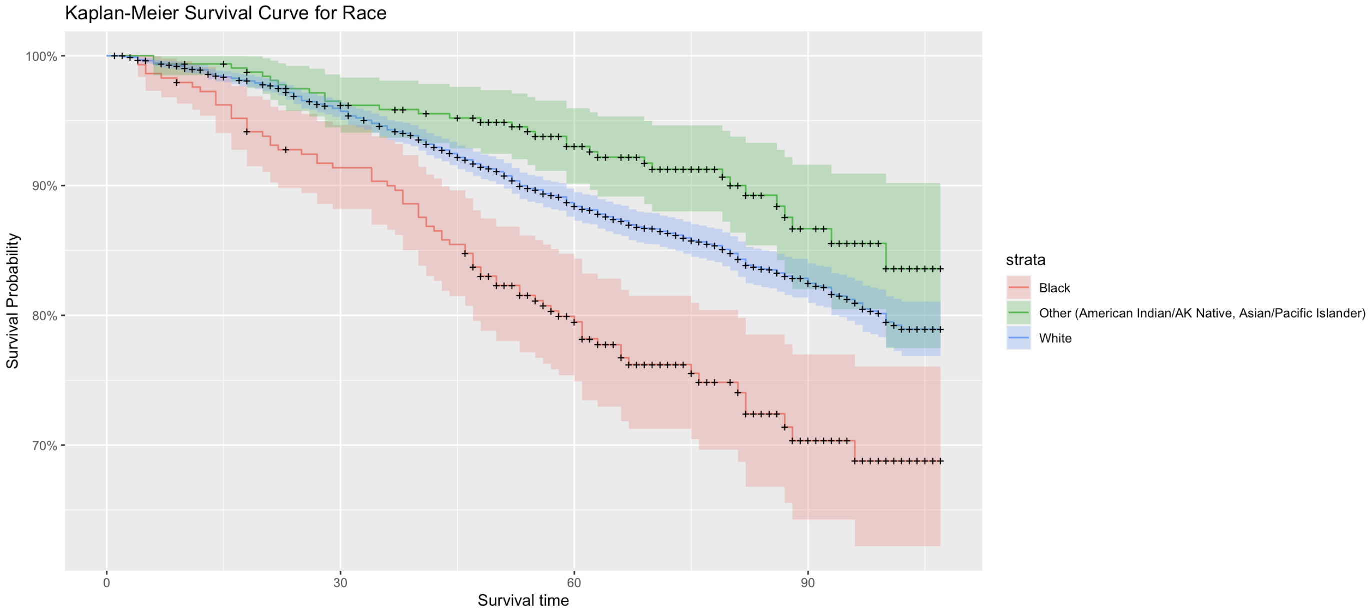
Here are the Kaplan-Meier plots generated from all the covariates in the dataset. In the study, all continuous variables are divided into four groups based on their quantiles to simplify the process and enhance the clarity of the plots.

Figure 4.1: The Kaplan–Meier curve for age



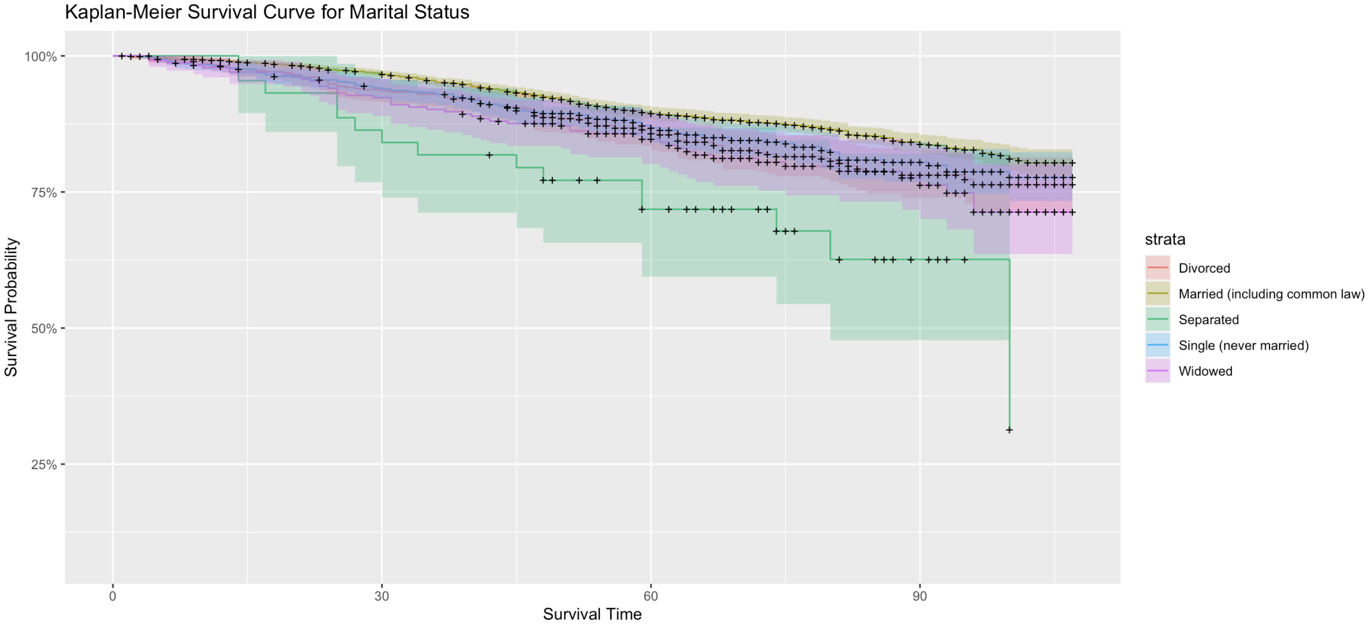
From the Figure 4.1, it can be observed that the survival probability and age is related negatively. For each curve, with the increase of age, the survival probability decreases. Moreover, the survival probability varies between each group. 'Q4' group (the purple line) represents the subjects with highest ages in the dataset, and its survival probability is obviously lower than the other groups. In terms of other three groups, although the difference between them is relatively small, it can still observe that the survival probability for groups with lower ages is higher than that for groups with larger ages. Additionally, '+' signs on the curve represents censored data. Therefore, it can be reasonably concluded that the survival probability is lower for elder people with breast cancer compared to young people with breast cancer.

Figure 4.2: The Kaplan–Meier curve for race



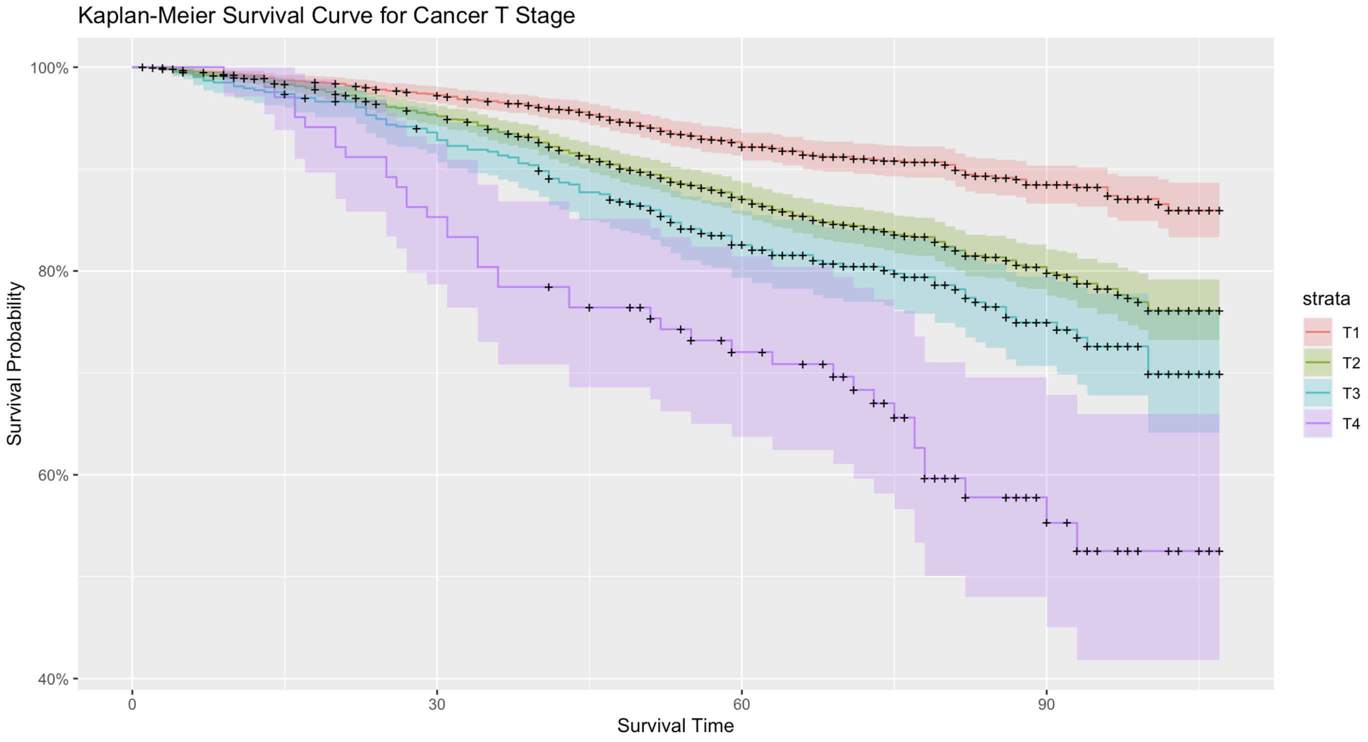
From the Figure 4.2, it is evident that the survival probability for black individuals is notably lower compared to other racial groups, specifically Others and Whites. Conversely, the group 'Other' exhibits the highest survival probability among all the groups depicted. Additionally, it's worth noting that the confidence interval for the black group appears to be broader in comparison to the intervals of other groups, indicating higher uncertainty in the estimation of survival probabilities within that demographic. On the other hand, the confidence interval for the white group is narrow among all the groups, which means that white group has the lowest uncertainty in the estimation of survival probabilities.

Figure 4.3: The Kaplan–Meier curve for marital status



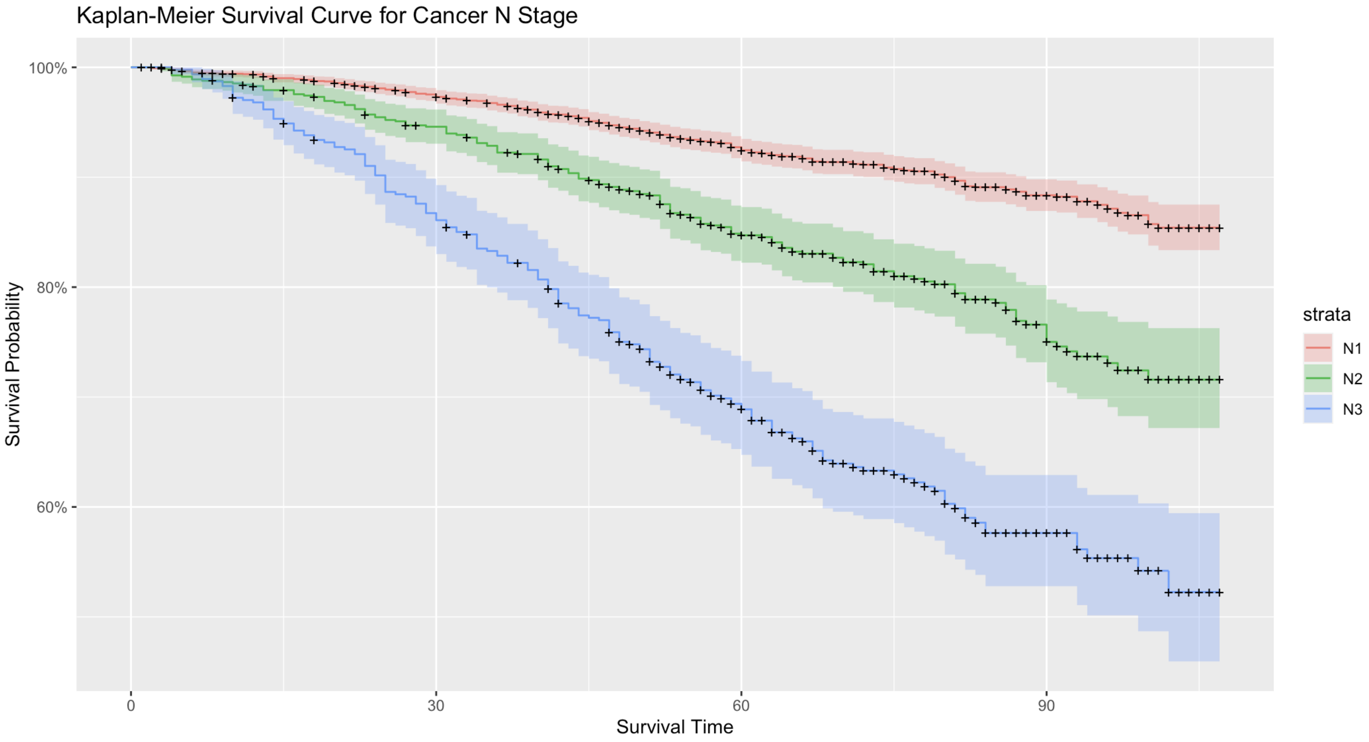
The Figure 4.3 displays survival probabilities over time across five marital statuses. Married individuals exhibit the highest and most stable survival probability, followed by single and separated categories which show moderate declines. The divorced group's survival probability decreases more noticeably, and the widowed group's curve declines the fastest, indicating the lowest survival probability over time. The shaded areas around each line represent the confidence intervals, and the 'plus' symbols indicate censored data, where participants left the study or were lost to follow-up. The trends suggest differences in survival based on marital status, with some groups consistently faring better or worse over the observed period.

Figure 4.4: The Kaplan–Meier curve for cancer T stage



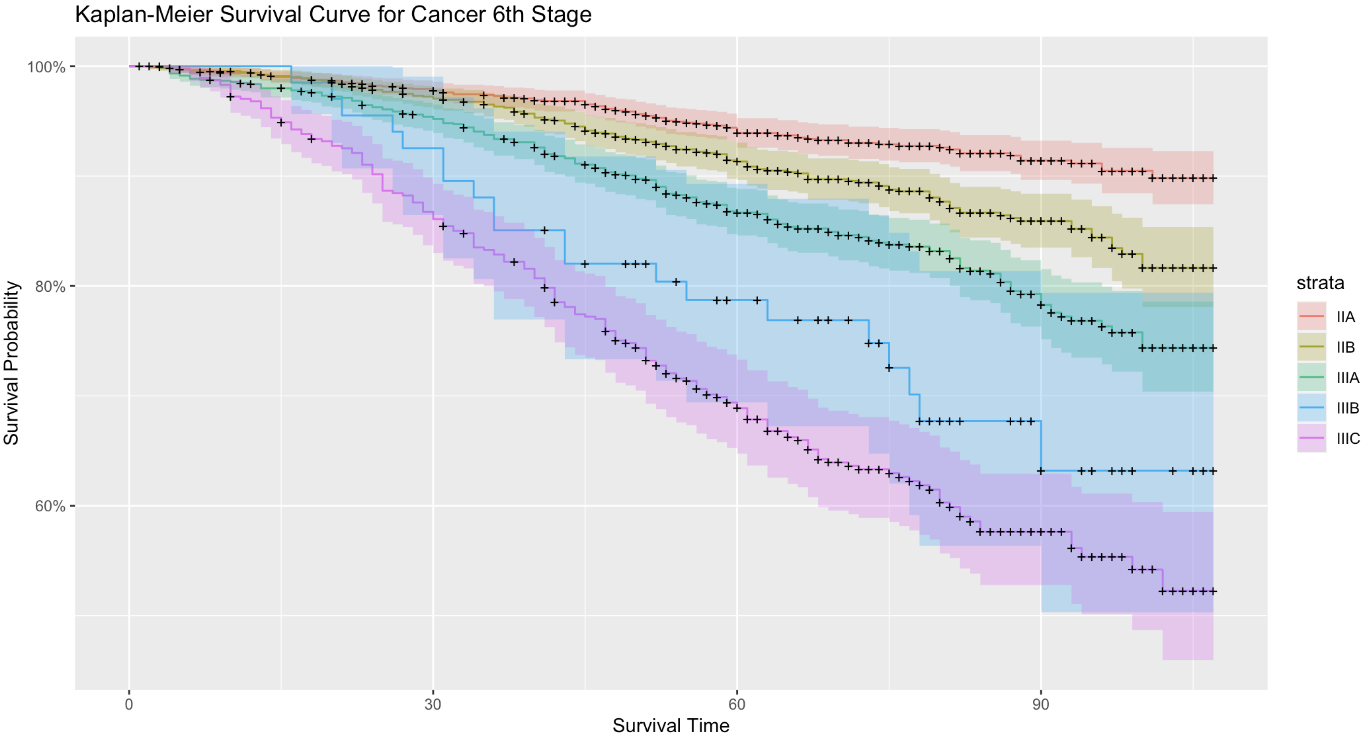
The Figure 4.4 represents survival probabilities for different cancer T stages over time. The 'T' in T stage refers to the size and extent of the main tumor. The four curves, each color-coded for stages T1 through T4, show that higher stages correspond to lower survival probabilities. T1 has the highest survival probability, indicating that patients with the smallest tumors fare better over time, while T4, representing the most advanced tumors, has the lowest survival probability. The plot's declining lines reflect decreasing survival probabilities as time progresses, with the shaded areas around each curve representing confidence intervals. The plus symbols indicate censored data points where patients were lost to follow-up or the study ended without an event occurring for those patients. The trend suggests that as the T stage increases, the survival probability decreases.

Figure 4.5: The Kaplan–Meier curve for cancer N stage



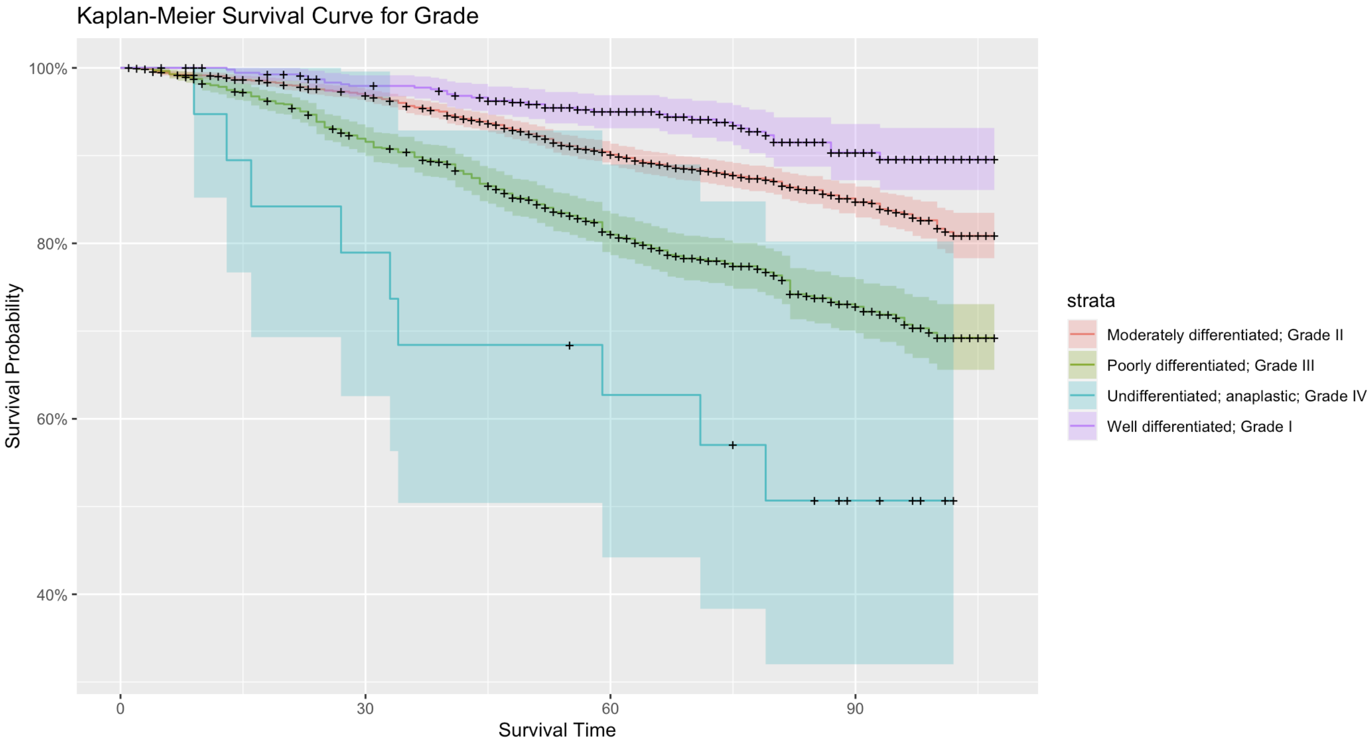
The Figure 4.5 presents a comparative survival analysis for patients across three nodal involvement levels: N1, N2, and N3. The red line represents N1, typically indicating minimal lymph node involvement, and it shows a relatively gentle slope downwards, indicating a higher survival probability over time. The green line for N2, marking intermediate lymph node involvement, reveals a steeper descent, suggestive of a quicker reduction in survival probability. The most pronounced decline is seen in the blue line for N3, which denotes extensive nodal involvement, correlating with the lowest survival rates across the time spectrum. The bands surrounding each line illustrate the confidence intervals, reflecting the statistical uncertainty of the survival estimates. Plus symbols along the survival curves indicate censored observations, representing patients who were lost to follow-up or whose study ended before an event occurred. The plot clearly demarcates the survival impact of increasing nodal involvement in cancer progression.

Figure 4.6: The Kaplan–Meier curve for cancer 6th stage



The Figure 4.6 here compares survival probabilities for various sub-stages of cancer stage 6: IIA, IIB, IIIA, IIIB, and IIIC. The plot indicates that patients with stage IIA cancer, represented by the lightest colored line at the top, have the highest survival probability. As the sub-stages progress towards IIIC, depicted by the purple line at the bottom, survival probabilities decrease notably. The plot shows a clear stratification of survival outcomes based on the sub-stage, with more advanced stages correlating with lower survival probabilities. Shaded areas indicate the confidence intervals for each sub-stage, offering a visual representation of the statistical uncertainty associated with the survival estimates. Plus signs on the curves represent censored data points for patients who were lost to follow-up or survived until the end of the study period.

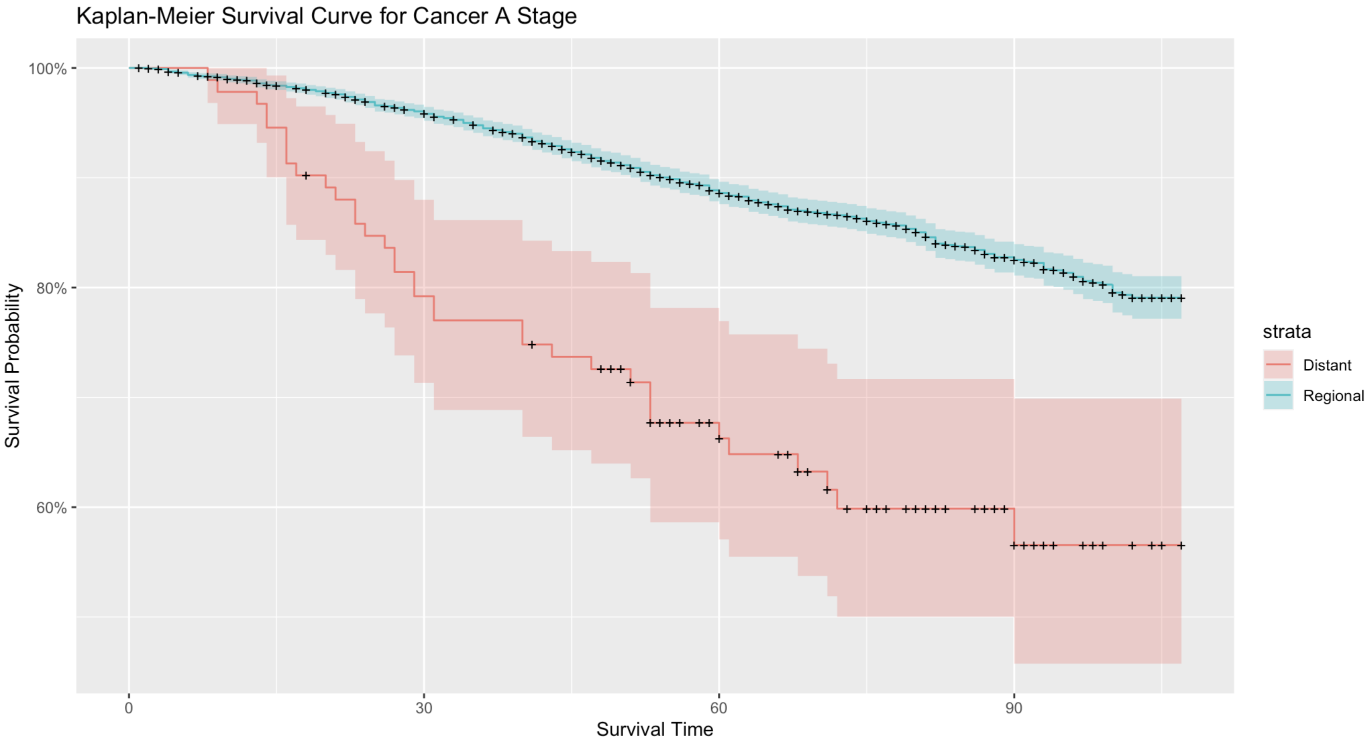
Figure 4.7: The Kaplan–Meier curve for grade



The Figure 4.7 for cancer grade demonstrates a clear stratification of patient survival based on cellular differentiation. Well-differentiated (Grade I) tumors, shown in purple, have the highest survival probabilities, reflecting their closer resemblance to normal cells. As the grade increases in severity to moderately differentiated (Grade

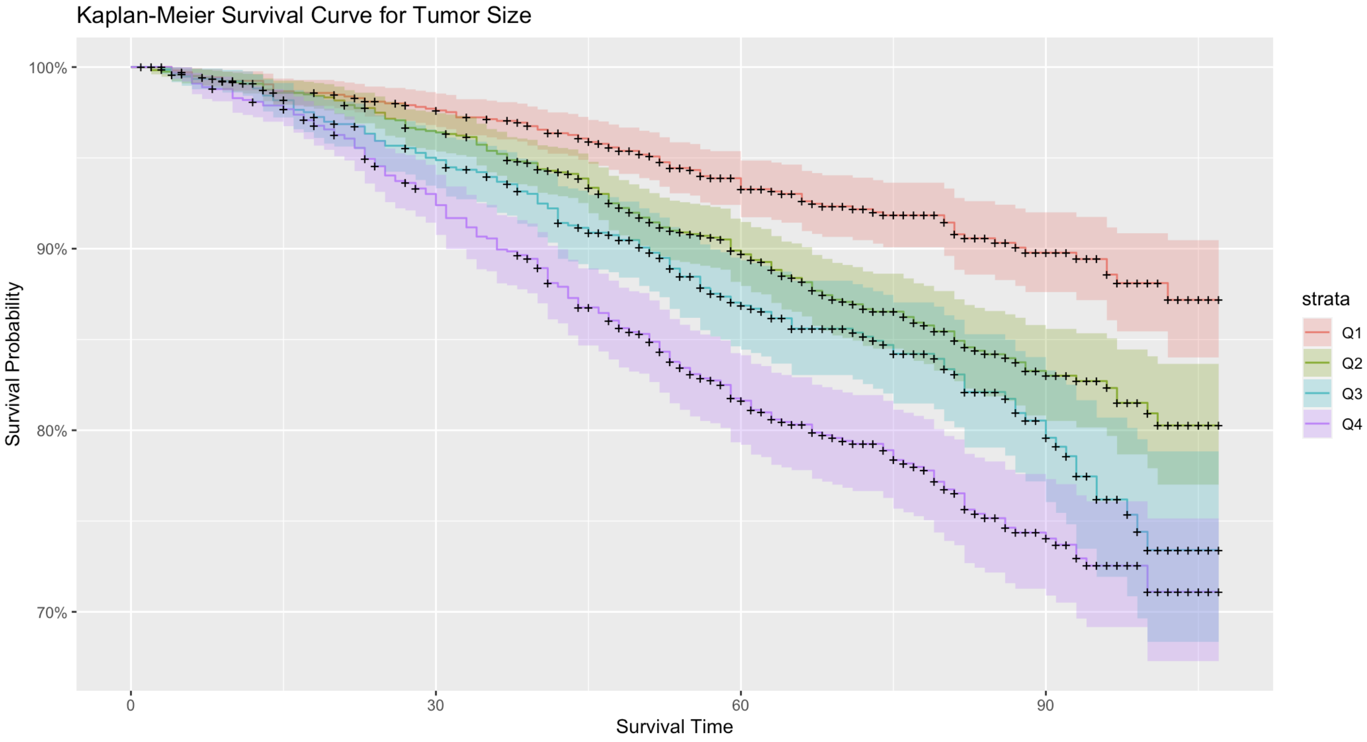
II in red), poorly differentiated (Grade III in green), and undifferentiated (Grade IV in light blue), survival probabilities correspondingly decrease. The steepest decline is seen with undifferentiated cancers, indicating the most aggressive behavior and poorest prognosis. The shaded regions denote confidence intervals, while plus signs mark censored data, signifying patients who did not have the event (e.g., death) occur during the study period.

Figure 4.8: The Kaplan–Meier curve for cancer A stage



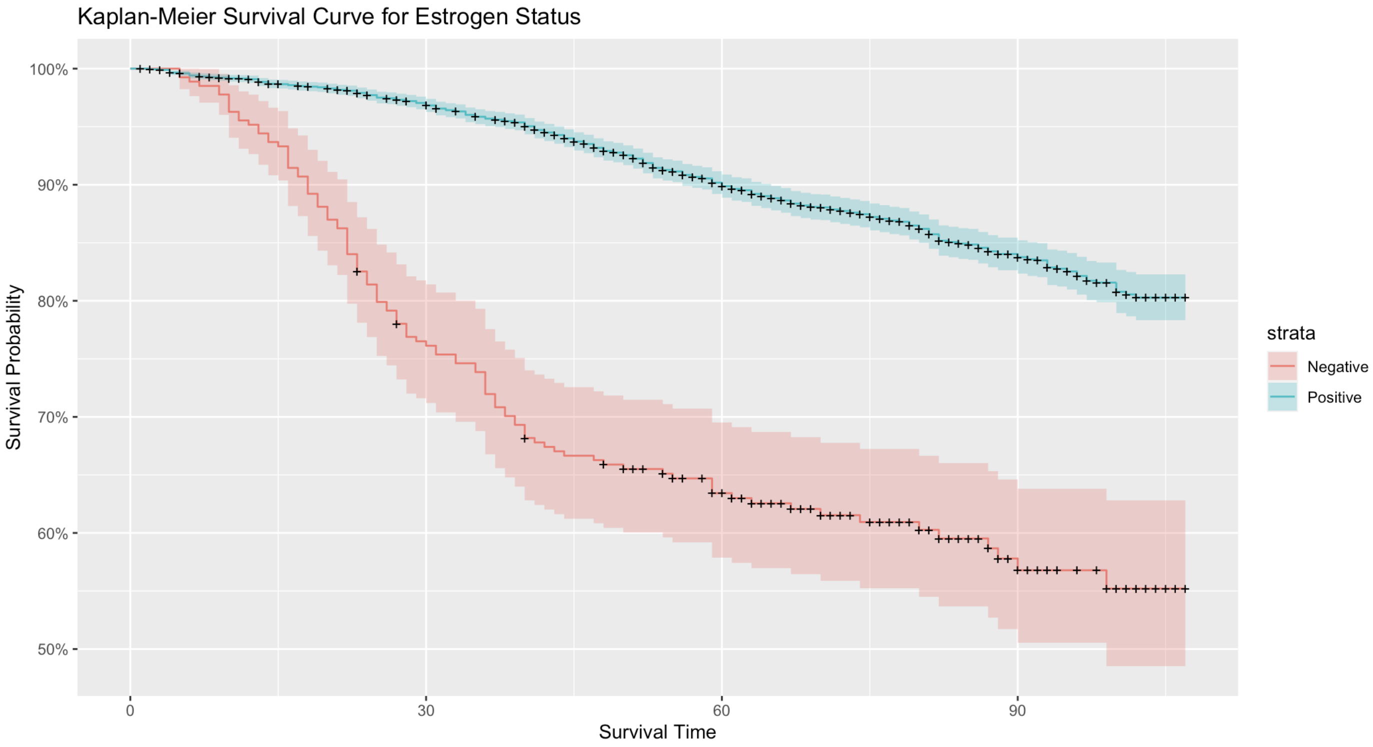
From Figure 4.8, two distinct cancer stages, Regional and Distant, are compared in terms of patient survival over time. The teal curve represents patients with Regional stage cancer, showing a higher survival probability that gently slopes downward, indicating a better prognosis for these patients. Conversely, the pink curve denotes the Distant stage cancer, which has a much steeper decline, reflecting a significantly lower survival probability over time and a worse prognosis. The plot's plus symbols indicate censored data points, where patients were lost to follow-up before the study's end or the event of interest occurred. The shaded areas represent the confidence intervals, highlighting the precision of the survival probability estimates.

Figure 4.9: The Kaplan–Meier curve for tumor size



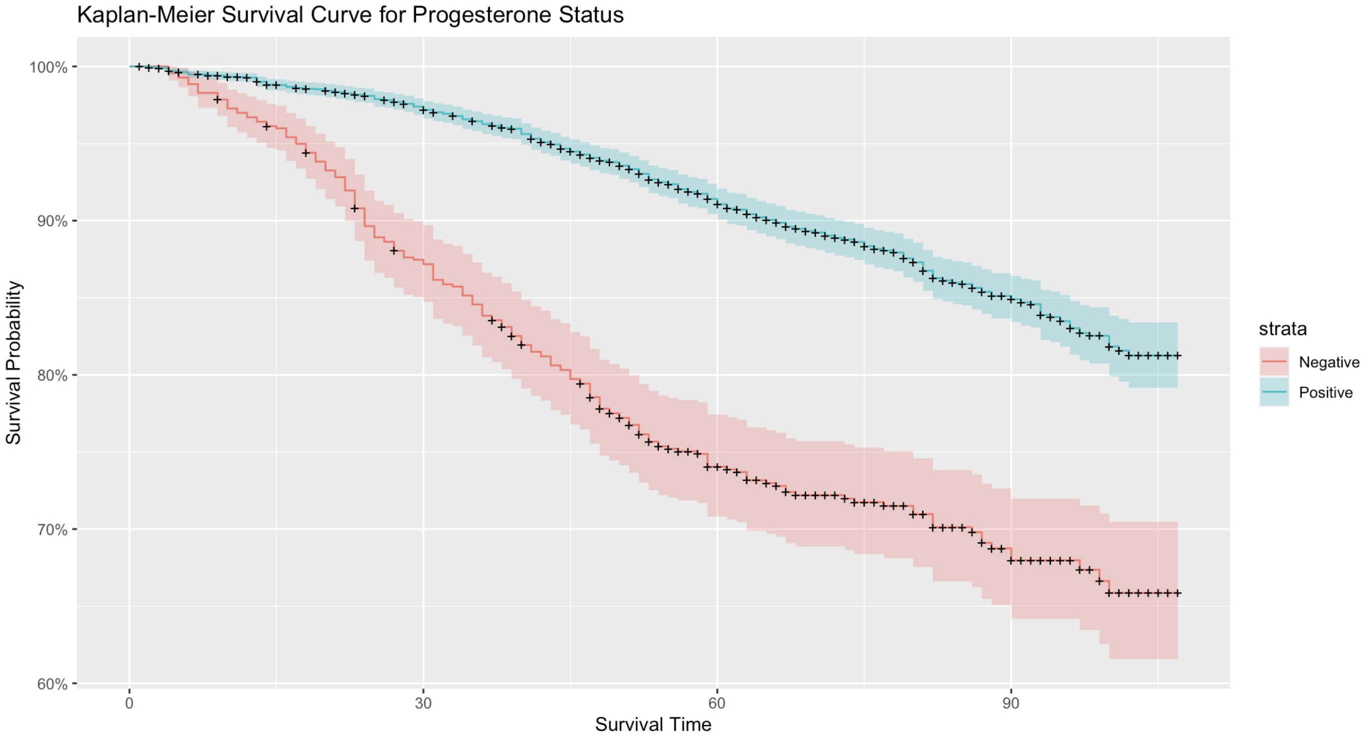
The Figure 4.9 compares the survival probabilities of cancer patients across four quantiles (Q1-Q4) of tumor size, suggesting tumor size was categorized into four equal groups based on patient distribution. The curve for Q1, representing the smallest tumors, shows the highest survival probability, with a gradual decrease over time. As tumor size increases through Q2 and Q3, there is a corresponding stepwise decrease in survival probabilities. The Q4 group, representing the largest tumors, has the lowest survival probability with the steepest decline. The shaded areas around each curve illustrate the confidence intervals, while plus symbols mark censored data points, indicating patients who were lost to follow-up or were still alive at the study's end. The plot reveals a clear trend: larger tumor sizes are associated with lower survival probabilities.

Figure 4.10: The Kaplan–Meier curve for estrogen status



The Figure 4.10 illustrates the survival probabilities of patients based on their estrogen status, with two distinct groups represented: estrogen-positive and estrogen-negative. The teal curve signifies patients who are estrogen-positive and shows a more favorable survival probability that declines slowly over time. In contrast, the pink curve indicates estrogen-negative patients, displaying a steeper decline in survival probability, suggesting a less favorable outcome. The shaded areas around each curve represent the confidence intervals, providing a visual measure of the estimate's precision, while the plus symbols indicate censored data, for patients who were lost to follow-up or did not experience the event by the study's end.

Figure 4.11: The Kaplan–Meier curve for progesterone status



This Figure 4.11 displays survival probabilities over time for patients with different progesterone receptor (PR) statuses, divided into PR-positive and PR-negative groups. The teal curve represents PR-positive patients, showing a higher survival probability with a more gradual decline over time. The pink curve, corresponding to PR-negative patients, exhibits a steeper decrease in survival probability, indicating a less favorable prognosis. The shaded bands around the curves indicate the confidence intervals, capturing the variability of the survival estimates. Plus signs along the curves denote censored data, representing patients who either left the study early or were alive at the end of the study period.

Figure 4.12: The Kaplan–Meier curve for regional node examined

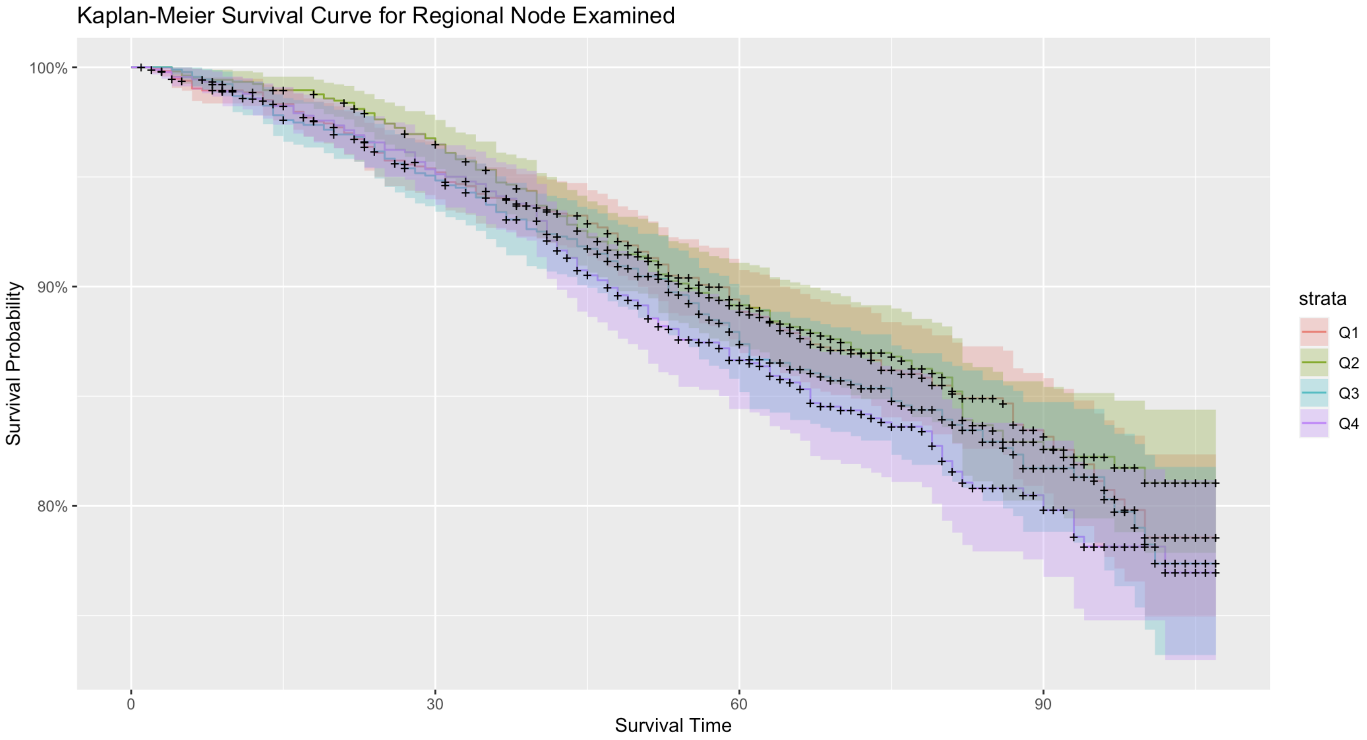
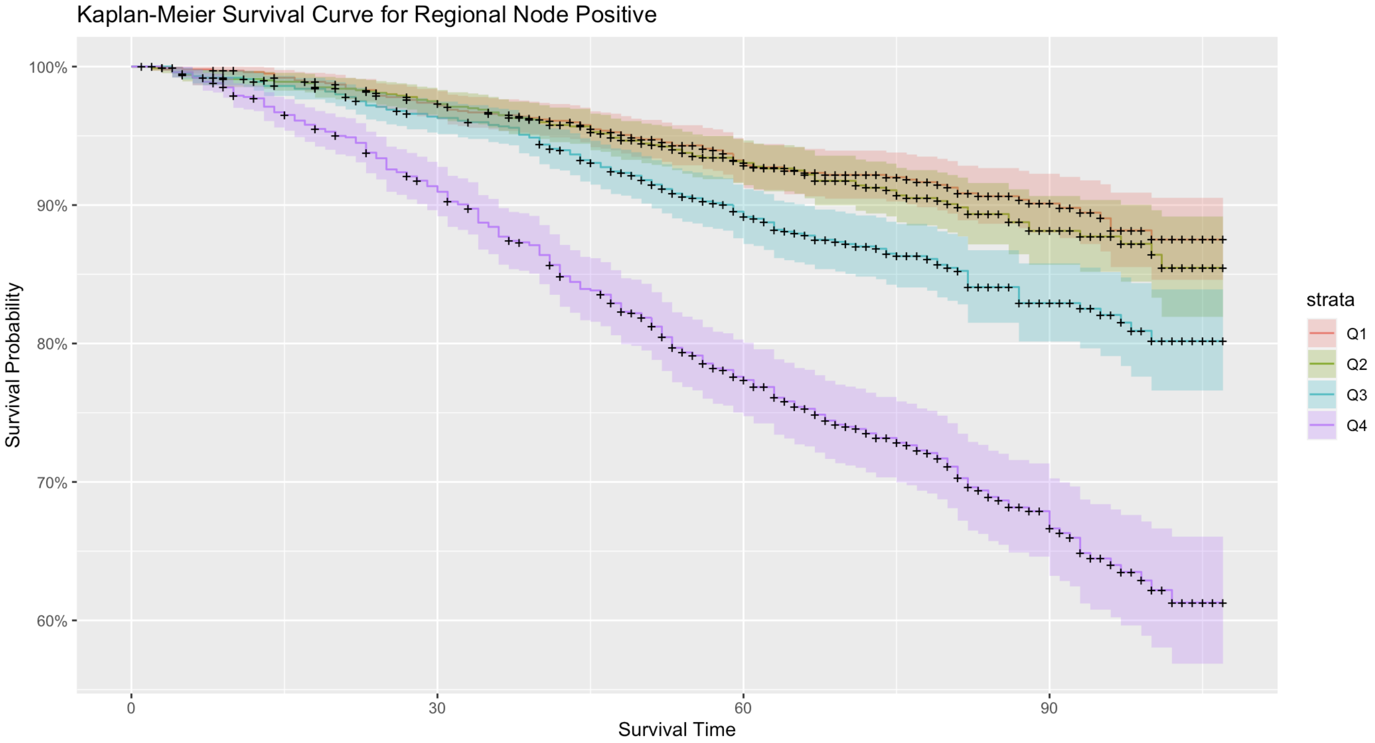


Figure 4.12 represents survival probabilities based on the number of regional lymph nodes examined in cancer patients, divided into four quantiles (Q1-Q4) to handle the continuous nature of the variable. The patients in Q1, likely having the fewest nodes examined, show the highest survival probability, and this probability decreases with each subsequent quantile. Q4, representing the patients with the most nodes examined, has the lowest survival probability and the steepest decline. This

suggests a potential association between the number of lymph nodes examined and patient survival outcomes. The plot's descending lines with plus symbols indicate censored data, while the surrounding shaded areas provide confidence intervals for each quantile's survival estimate.

Figure 4.13: The Kaplan–Meier curve for regional node positive



The Figure 4.13 depicted here illustrates the survival probabilities for patients with regional lymph node involvement, categorized into four groups (Q1-Q4) based on quantiles of the number of positive nodes. Q1, with the fewest positive nodes, displays the highest survival probabilities, starting near 100% and declining gently over time. Survival probabilities progressively decrease with higher quantiles, indicating more lymph node involvement. Q4 shows the lowest survival probability, with the most significant decrease, suggestive of a higher disease burden. The plot's descending trend lines, along with plus symbols for censored data and shaded confidence intervals, underscore the impact of lymph node involvement on survival outcomes.

4.2 Logrank Test

The following table is the summary of the logrank test results for each variable in the dataset:

Table 4.1: Logrank Test results

Variable name	P-value
Age	5×10^{-6}
Race	2×10^{-7}
Marital Status	2×10^{-6}
T Stage	$< 2.2 \times 10^{-16}$
N Stage	$< 2.2 \times 10^{-16}$
6th Stage	$< 2.2 \times 10^{-16}$
Grade	$< 2.2 \times 10^{-16}$
A Stage	5×10^{-12}
Tumor Size	0.3
Estrogen Status	$< 2.2 \times 10^{-16}$
Progesterone Status	$< 2.2 \times 10^{-16}$
Regional Node Examined	0.3
Regional Node Positive	$< 2.2 \times 10^{-16}$

From the results for each variable, since the p values for ’tumor size’ and ’regional node examined’ variables are larger than the given significant level ((α)= 0.05), it can be concluded that we do not have enough evidence to reject the null hypothesis. In other words, there are no distinctions between the populations in terms of the likelihood of an event for ’tumor size’ and ’regional node examined’ variables.

Moreover, since the p values for other variables in the dataset are smaller than the given significant level, we have enough evidence to reject the null hypothesis, which means that there are distinctions between the populations in terms of the likelihood of an event.

4.3 Cox Proportional Hazards Model

In order to select the most appropriate model for the dataset, the step() function is utilized. The function with the lower AIC value is selected as the final model for the dataset.

The final model is summarized as :

Variable	Coefficient	Hazard Ratio	95% Confidence Interval
age	0.021	1.021	(1.012, 1.030)
raceOther	-0.825	0.438	(0.290, 0.663)
raceWhite	-0.434	0.648	(0.506, 0.830)
cancer_TstageT2	-0.390	1.478	(1.213, 1.800)
cancer_TstageT3	0.516	1.674	(1.304, 2.149)
cancer_TstageT4	0.948	2.576	(1.794, 3.675)
cancer_NstageN2	0.374	1.453	(1.250, 1.917)
cancer_NstageN3	0.566	1.762	(1.241, 2.501)
Grade I	-0.452	0.637	(0.455, 0.890)
Grade III	0.352	1.421	(1.193, 1.691)
Grade IV	1.074	2.926	(1.491, 5.742)
Estrogen_StatusPositive	-0.650	0.522	(0.401, 0.680)
Progesterone_StatusPositive	-0.493	0.611	(0.496, 0.753)
regional_node_examined	0.036	0.967	(0.955, 0.979)
regional_node_positive	0.063	1.065	(1.042, 1.089)

Table 4.2: Coefficient and Hazard Ratio Estimates from Cox PH Model

This table displays the results from the final Cox proportional hazards model.

Variable: The predictor or feature in the model.

Coefficient: A value representing the log hazard ratio for the corresponding variable. A positive coefficient suggests that as the value of the variable increases, the event hazard increases, while a negative coefficient suggests a decrease in hazard with an increase in the variable.

Hazard Ratio: This is the exponentiated coefficient, indicating the factor by which the hazard (risk of the event occurring) is multiplied for a one-unit increase in the variable. A hazard ratio above 1 indicates an increased hazard, while a hazard ratio below 1 indicates a decreased hazard.

95% Confidence Interval: To display the uncertainty of the hazard ratio,the 95% Confidence Interval of the hazard ratio is included in the table.

Here's a brief interpretation of the results:

- **age**: For each one-year increase in age, the hazard increases by a factor of about 1.021, which means older individuals have a slightly higher risk of the event happening.

- **raceOther**: Being of a race classified as "Other" (American Indian/AK Native, Asian/Pacific Islander) is associated with about a 56.2% decrease in the hazard compared to the baseline race (Black), since the hazard ratio is less than 1.

- **raceWhite**: Being white is associated with a 35.2% decrease in the hazard compared to the baseline race.

- **cancer_Tstage**: The various stages of cancer T-stage have different effects. For instance, T2 is associated with a 47.8% increase in hazard, T3 with a 67.4% increase, and T4 with a 157.6% increase, indicating progressively higher risks with advancing T-stage. The baseline level used for comparison is T1 stage.

- **cancer_NstageN2 and cancer_NstageN3**: Higher N-stages indicate a higher hazard, with N3 being associated with a 76.2% increase in hazard. The baseline level used for comparison is N1 stage.

- **Grade**: Different cancer grades show different risks. Poorly differentiated (Grade III) increases hazard by 42.1%, while undifferentiated/anaplastic (Grade IV) nearly triples the hazard. In contrast, well-differentiated (Grade I) decreases the hazard by 36.3%. The baseline level used for comparison is Grade II.

- **Estrogen_StatusPositive**: A positive estrogen status is associated with a decrease in hazard by about 47.7%. The baseline level used for comparison is negative estrogen status.

- **Progesterone_StatusPositive**: A positive progesterone status is associated with a decrease in hazard by about 39.0%. The baseline level used for comparison is negative progesterone status.

- **Regional_node_examined and Regional_node_positive**:The hazard ratio for examined regional node is 1.036199, which indicates that for each additional examined regional node , there is a 3.62% increase in the hazard, and the hazard ratio for positive regional node is 1.06545, so for each additional positive regional node, there is 6.55% increase in the hazard.

4.4 Schoenfeld Residual Diagnosis

In order to assess if the proportional hazards assumption is met for the final model selected by step() function, the Schoenfeld residuals and score test is performed. Here are the results for the test:

Variable	P Value
Age	0.76
Race	0.57
Cancer_Tstage	0.97
Cancer_Nstage	0.38
Grade	0.55
Estrogen_Status	3.0×10^{-8}
Progesterone_Status	8.4×10^{-9}
Regional_node_examined	0.87
Regional_node_positive	0.79

Table 4.3: The Schoenfeld score test results

From the result, it can be observed that the covariates except 'Estrogen Status' and 'Progesterone Status' have the p-value larger than the significant level ($(\alpha) = 0.05$). Therefore, we do not have enough evidence to reject null hypothesis, which can be concluded that there is no violation of the proportional hazards assumption for these covariates. However, since the p-values for 'Estrogen Status' and 'Progesterone Status' are very small, which means that there is a violation of the proportional hazards assumption for these two covariates. With the effects of the significant covariates, the global test result has the p-value much smaller than the significant level, so the violation of the proportional hazards assumption exists for the final model.

To further explore the result of the Schoenfeld individual test of the final model, the Schoenfeld individual test plots are utilized. Here are the plots generated from the final model:

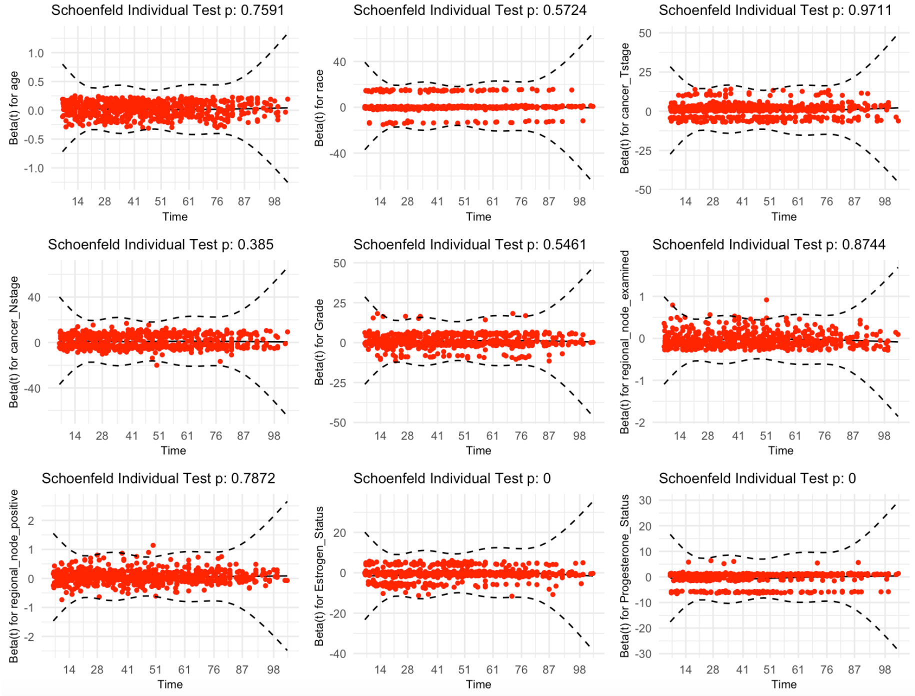


Figure 4.14: The Schoenfeld individual test plots

If the Schoenfeld residuals plots exhibit a horizontal or random pattern, the proportional hazards assumption is not violated. Conversely, if the Schoenfeld residuals plots exhibit a clear and consistent pattern, the proportional hazards assumption is violated [24].

It can be observed from the plots generated from the final model, covariates expect 'Estrogen Status' and 'Progesterone Status' manifest random or horizontal patterns, so they do not violate the proportional hazards assumption. On the other hand, the patterns for 'Estrogen Status' and 'Progesterone Status' are curvy with consistent patterns, so they violate the proportional hazards assumption.

4.5 Martingale Residuals

The martingale residuals plot was an important tool in accessing the goodness-of-fit. If the scatter plot does not show any obvious patterns or systematic deviations from the horizontal line at zero, this typically suggests that the Cox proportional hazards model may have a reasonable goodness of fit to the data. A lack of pattern means that the model is not systematically over or under predicting the hazard at different times, which is an indication that the functional form of the covariates in the model might be appropriate.

Here is the martingale residuals plot generated from the continuous variables in the dataset (age, tumor size, regional node examined, and regional node positive) of this research:

Figure 4.15: The Martingale Residuals Plot for Age

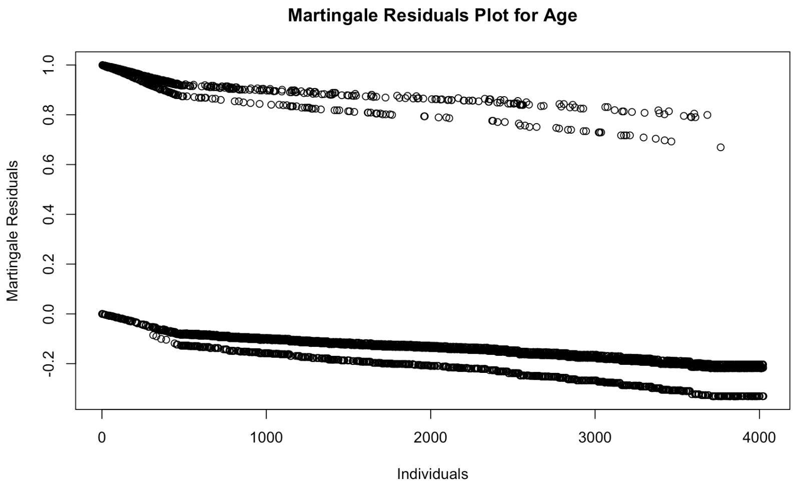


Figure 4.16: The Martingale Residuals Plot for Tumor Size

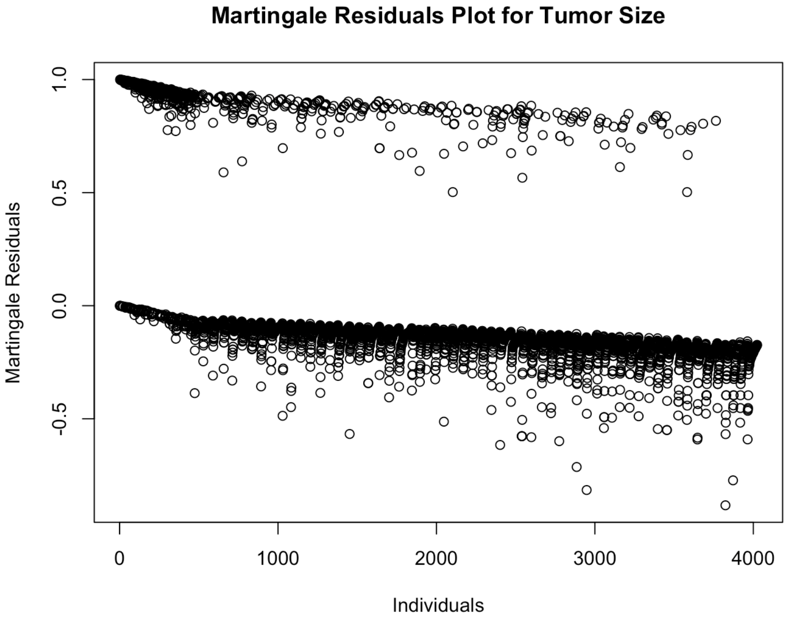


Figure 4.17: The Martingale Residuals Plot for Regional Node Examined

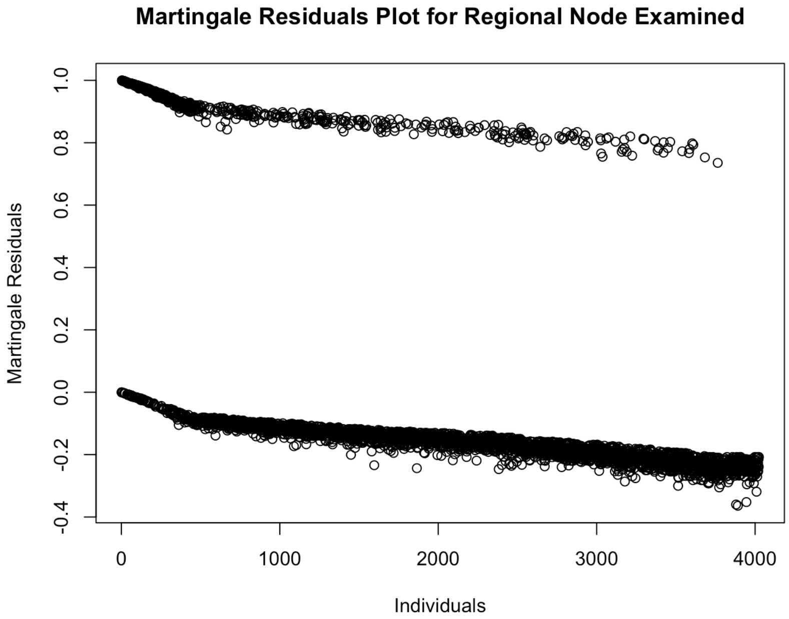
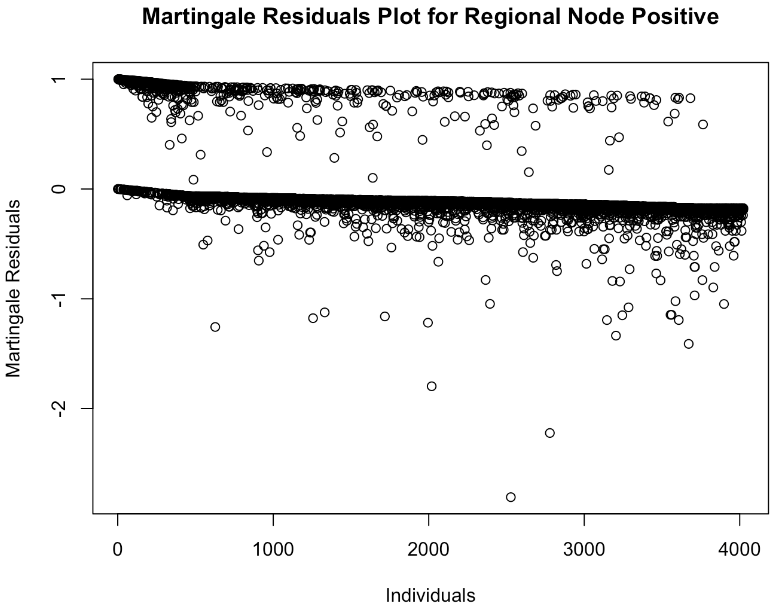


Figure 4.18: The Martingale Residuals Plot for Regional Node Positive



From the above plots, it can be observed that no obvious pattern is shown in the graph. This lack of pattern suggests that the model may not be missing any major non-linear effects of the covariates. Moreover, the residuals cluster around the zero line, so the model reasonably fitted all the continuous data. Additionally, although there are some potential outliers shown in the graph, it cannot be ensured if the outliers exist in the data due to martingale residuals' inherent property of lack of symmetry.

The model's goodness of fit cannot be affirmed from the martingale residuals plot alone, but it can be concluded that there are no major evident issues with the model's fit to the data.

Chapter 5

Discussion

5.1 Comparing non-parametric models and the Cox proportional hazards model

Non-parametric models and the Cox proportional hazards model are commonly used in survival analysis due to their appropriate approach to censored data, which is the most unique characteristic in survival analysis. Both advantages and disadvantages exist in these models.

The most common non-parametric technique for modeling the survival function is the Kaplan-Meier estimate. The Kaplan-Meier estimate has benefits in flexibility, and the model complexities grow with the increase of observation sizes. However, there are two major disadvantages exist. Firstly, it is hard for covariates to be incorporated into the model. In other words, describing the differences between individuals in their survival functions is challenging. To deal with the problem, one approach is to compare different fitting models using different subpopulations, which is infeasible as the sample size becomes larger. Secondly, the survival functions are not smooth. Although it is possible that the functions become more smooth as the sample sizes increase, the models built from small sample sizes are not smooth and are likely to generate unrealistic properties [25].

From the result of my research, the small sample size with only around 4000 causes some limitations in the research. Firstly, it can be observed that the Kaplan-Meier curves are not smooth all the time. The sample size limits the interpretation and generalizability of the findings. Moreover, for the log-rank test, there exists results that have higher observation values than the actual observations, such as the Black racial group, which suggests potential disparities in survival outcomes. However, caution should be exercised in drawing definitive conclusions, as the relatively small sample size may introduce uncertainty and variability in the estimates. Further research with larger and more diverse cohorts is warranted to corroborate these findings and provide more robust insights into the impact of race on survival outcomes in this population.

With the advantage of avoiding reliance on uncertain assumptions, such as constant hazard and linearly increasing hazard, of the Cox Proportional hazard model, it is an ideal statistical method to construct a regression model for censored data. However, the Cox Proportional hazard model possesses some drawbacks. Firstly, the assumption of proportional hazards is often unrealistic, posing challenges in its verification and potentially resulting in significant biases. Since the Cox Proportional hazard model assumes that the hazard functions for different groups are proportional over time, so the model's results may be biased or inaccurate if this assumption is violated. Secondly, the Cox model exhibits lower statistical efficiency compared to models like Weibull models, which utilize the complete survival time information. Moreover, hazard ratios provided by the Cox model can sometimes be challenging to interpret, especially when covariates are categorical or time-dependent [26].

5.2 Potential Problems in the final Cox Proportional Hazard Model

The final Cox proportional hazard model includes the covariates of age, race, cancer T stage, cancer N stage, grade, estrogen status, progesterone status, examined regional node, and positive regional node. As mentioned in the test of proportional hazard assumptions, the covariates of Estrogen Status and Progesterone Status have the p values smaller than the significant level of 0.05. This suggests that the proportional hazards assumption for covariates is violated. Therefore, the whole model violates the proportional hazards assumption.

One of the reasons that causes this violation is the binary covariate's interaction with other variables in the model, leading to time-dependent effects on the hazard ratio over the observation period [27]. Estrogen Status and Progesterone Status are binary covariates with two levels of positive and negative, so it is possible that their effects on survival change over time, especially in the context of hormone-related treatments or interventions. Additionally, the presence of unmeasured confounding factors related to hormonal fluctuations or treatment responses may further contribute to deviations from proportional hazards. Therefore, careful consideration of these factors and potentially incorporating time-varying covariates or stratification techniques may be necessary to address the violation of the proportional hazards assumption in Cox regression modeling.

Stratified Cox model is a plausible way to solve the problem. The Stratified Cox model is an adaptation of the Cox proportional hazards (PH) model, enabling the management of predictors that do not adhere to the PH assumption through the technique of stratification. This approach involves dividing the dataset into distinct groups based on the stratifying variable, allowing for separate baseline hazard functions within each stratum while still estimating the effects of other covariates consistently across all strata. By stratifying on variables that violate the PH assumption, such as categorical predictors or time-varying covariates, the Stratified Cox model provides a flexible and robust framework for survival analysis, enhancing the accuracy and reliability of the results [28].

An alternative model that can be employed is Accelerated failure time(AFT) model. The accelerated failure time (AFT) model stands as a parametric alternative to the more prevalent proportional hazards models. While the proportional hazards model posits that a covariate impacts the hazard by a multiplicative constant, the AFT model posits that a covariate affects the timeline of a disease's progression by a constant factor, either hastening or delaying the event of interest. It's referred to as "accelerated" because the model essentially estimates how much the covariates speed up (accelerate) or slow down (decelerate) the time to the event [29]. Unlike the Cox model, which models the hazard rate, the AFT model directly estimates the effect of covariates on the survival time. It assumes that covariates accelerate or decelerate the life process by a constant factor. The AFT model focuses on survival time as the outcome of interest, rather than the hazard rate. This focus on time-to-event rather than the rate of event occurrence makes the AFT model suitable for situations where the hazards are not proportional. Therefore, Accelerated failure time(AFT) model is an ideal alternative model when the proportional hazards assumptions are not held by the Cox proportional hazards model[30].

Chapter 6

Reference

[1] British Medical Journal. (n.d.). *Survival analysis.* https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/12-survival-analysis

[2] T.G. Clark, M.J. Bradurn, S.B. Love D.G. Altman. (July 15, 2003). *Survival Analysis Part I: Basic concepts and first analyses.* British Journal of Cancer. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2394262/

[3] EUI. (February 7, 2024). *Zenodo open data repository* (CERN). https://www.eui.eu /Research/Library/ResearchGuides/Economics/Statistics/DataPortal/Zenodo

[4] M. G. Akritas. (2004). *Nonparametric Survival Analysis.* [PDF]. Statistical Science.

[5] S. Taylor. (n.d.). *Nonparametric Tests.* CFI. https://corporatefinanceinstitute.com /resources/data-science/nonparametric-tests/

[6] Semiparametric model. (June 17, 2021). In *Wikipedia.* https://en.wikipedia.org /wiki/Semiparametric$_{m}odel$

[7] Censoring(statistics). (November 1, 2023). In *Wikipedia.* https://en.wikipedia.org /wiki/Censoring$_{(}statistics)$

[8] Reid, M. (2022). *What is censored data?* Github. https://github.com/MatthewReid854/reliability/blob/master/docs /What%20is%20censored%20data.rst

[9] University of Washington. (n.d.). *Lecture 17* [PDF]. BIOST 515. https://courses.washington.edu/b515/l17.pdf

[10] Survival analysis. (March 25, 2024). In *Wikipedia.* https://en.wikipedia.org /wiki/Survival$_a nalysis$

[11] Kaplan-Meier estimator. (April 3, 2024). In *Wikipedia.* https://en.wikipedia.org /wiki/Kaplan%E2%80%93Meier$_e stimator$

[12] Goel, M. K., Khanna, P., Kishore, J. (2010). *Understanding survival analysis: Kaplan-Meier estimate.* International journal of Ayurveda research, 1(4), 274–278. https://doi.org/10.4103/0974-7788.76794

[13] DATAtab Team. (2024). *Kaplan Meier Curve.* https://datatab.net/tutorial/kaplan-meier-curve

[14] Logrank test. (February 8, 2024). In *Wikipedia.* https://en.wikipedia.org /wiki/Logrank$_t est$

[15] Bland, J. M., Altman, D. G. (2004). *The logrank test.* BMJ (Clinical research ed.), 328(7447), 1073. https://doi.org/10.1136/bmj.328.7447.1073

[16] Dawson, D.V., Blanchette, D.R. Pihlstrom, B.L. (2021). *13 - Application of Biostatistics in Dental Public Health.* Burt and Eklund's Dentistry, Dental Practice, and the Community (Seventh Edition). https://www.sciencedirect.com /science/article/abs/pii/B9780323554848000137

[17] Johnson, L.L, Shih, J.H. (2007). *CHAPTER 20 - An Introduction to Survival*

Analysis. Principles and Practice of Clinical Research (Second Edition). https://www.sciencedirect.com/science/article/abs/pii/B9780123694409500244

[18] Halabi, S., Dutta, S., Wu, Y., Liu, A. (2020). *Score and deviance residuals based on the full likelihood approach in survival analysis.* Pharmaceutical statistics, 19(6), 940–954. https://doi.org/10.1002/pst.2047

[19] Akaike information criterion. (February 2, 2024). In *Wikipedia.* https://en.wikipedia.org/wiki/Akaike_information_criterion

[20] Bevans, R. (March 26, 2020). *Akaike Information Criterion — When How to Use It (Example).* Scribbr. https://www.scribbr.com/statistics/akaike-information-criterion/

[21] Zajic, A. (November 29, 2022). *What Is Akaike Information Criterion (AIC)?* Builtin. https://builtin.com/data-science/what-is-aic

[22] Time Series Reasoning. (n.d.). *Schoenfeld residuals.* https://timeseriesreasoning.com/contents/schoenfeld-residuals/

[23] Breheny, P. (n.d.). *Residuals and model diagnostics* [PowerPoint slides]. BIOS 7210: Survival Data Analysis. The University of Iowa.

[24] Information Builders. (n.d.). *Topic 40.* https://infocenter.informationbuilders.com /wf8007/index.jsp?topic=%2Fpubdocs%2FRStat15%2Fsource%2Ftopic40.htm

[25] A. (November 25, 2018). *When Should You Use Non-Parametric, Parametric, and Semi-Parametric Survival Analysis.* Boostedml. https://boostedml.com/2018/11/when-should-you-use-non-parametric-parametric-and-semi-parametric-survival-analysis.html

[26] H.T.H. (n.d.). *What are the pros and cons of using Cox regression?* Quora. https://www.quora.com/What-are-the-pros-and-cons-of-using-Cox-regression

[27] Austin P. C. (2018). *Statistical power to detect violation of the proportional hazards assumption when using the Cox regression model.* Journal of statistical computation and simulation, 88(3), 533–552. https://doi.org/10.1080/00949655.2017.1397151

[28] Guo, W. (2010). *Chapter 5:The Stratified Cox Procedure.* Survival Analysis: A Self-Learning Text Book. https://web.njit.edu/ wguo/Math%20659_2010/Chapters %205%20and%206%20%20from%20%5BSurvival %20Analysis_A%20Self%20Learning%20Text_Book%5D.pdf

[29] Accelerated failure time model. (March 23, 2024). In *Wikipedia.* https://en.wikipedia.org/wiki/Accelerated_failure_time_model

[30] Swindell W. R. (2009). *Accelerated failure time models provide a useful statistical framework for aging research.* Experimental gerontology, 44(3), 190–200. https://doi.org/10.1016/j.exger.2008.10.005

[31] JMP. (n.d.). *Chi-Square Test of Independence.* https://www.jmp.com/en_au/statistics-knowledge-portal/chi-square-test/chi-square-test-of-independence.html

[32] Ford, C. (May 25, 2023). *The Wilcoxon Rank Sum Test.* University of Virginia. https://library.virginia.edu/data/articles/the-wilcoxon-rank-sum-test

Bibliography

Appendix A

Chi-square Test

The Chi-square statistic(χ^2) serves as a non-parametric tool, free from distribution assumptions, aimed at examining disparities between groups when the dependent variable is assessed categorically at a nominal level [31]. Specifically, the research utilized the Pearson's chi-square test for independence.

In the research, the chi-square for independence is performed between all categorical variables(race, marital status, T stage, N stage, grade, A stage, estrogen status, progesterone status) in the dataset and the status to test if the variable is associated with subjects' status.

The significant level(α) is 0.05.

Null hypothesis(H0): The variable and status are independent.

Alternative hypothesis(HA):The variable and status are not independent.

The following table is the summary of the results.

Table A.1: Chi-square Test Results

Variable name	Levels	P-value	Proportion
Race	Black, Other , White	8.411×10^{-7}	We have strong evidence to reject the null hypothesis.
Marital Status	28.264	1.103×10^{-5}	We have strong evidence to reject the null hypothesis.
T Stage	103.48	$< 2.2 \times 10^{-16}$	We have strong evidence to reject the null hypothesis.
N Stage	269.93	$< 2.2 \times 10^{-16}$	We have strong evidence to reject the null hypothesis.
Grade	112.56	$< 2.2 \times 10^{-16}$	We have strong evidence to reject the null hypothesis.
A Stage	35.765	2.226×10^{-9}	We have strong evidence to reject the null hypothesis.
Estrogen Status	135.16	$< 2.2 \times 10^{-16}$	We have strong evidence to reject the null hypothesis.
Progesterone Status	124.89	$< 2.2 \times 10^{-16}$	We have strong evidence to reject the null hypothesis.

All chi-square test results for the categorical variables were statistically significant ($p < 0.05$), indicating a significant association between each categorical variable and subject's status at the end of the study.

Appendix B

Wilcoxon Rank Sum Test

The Wilcoxon rank sum test is a non-parametric test used to test if two groups have the same distribution, specifically to compare whether they have the same median. The Wilcoxon rank sum test does not assume the variables are normally distributed, and it assumes the variables are independent from each other [32]. Therefore, the Wilcoxon rank sum test is ideal to assess the possible relationship between continuous variables and the variable of interest.

In the research, the Wilcoxon rank sum test is performed between all continuous variables (age, tumor size, regional node examined, regional node positive) in the dataset and the status of subjects at the end of the study.

The significant level (α) is 0.05.

Null hypothesis (H0): There is no difference between the distributions of the variable and status.

Alternative hypothesis (HA): There is difference between the distributions of the variable and status.

The following table is the summary of the results.

Table B.1: Wilcoxon Rank Sum Test results

Variable name	Test statistics	P-value	Result
Age	944505	7.34×10^{-5}	We have strong evidence to reject the null hypothesis.
Tumor Size	813236	$< 2.2 \times 10^{-16}$	We have strong evidence to reject the null hypothesis.
Regional Node Examined	1001152	0.0673	We do nothave strong evidence to reject the null hypothesis.
Regional Node Positive	692562	$< 2.2 \times 10^{-16}$	We have strong evidence to reject the null hypothesis.

From the results, it can be concluded that given the significant level equals 0.05, we have enough evidence to infer that the distributions of age, tumor size, and regional node positive are different from the distribution of the status. Moreover, because the p value of regional node examined is larger than 0.05, we do not have enough evidence to conclude that the distributions of regional node examined and the status are different.