Regression Analysis for Time-to-Event Data

Jasmine Kaur

MSc in Data Science & Statistics
The University of Bath
September 2023

Signed: Jasmine Kaur

Regression Analysis for Time-to-Event Data

submitted by

Jasmine Kaur

for the degree of MSc in Data Science & Statistics of the

The University of Bath

September 2023

COPYRIGHT

DECLARATION

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of MSc in Data Science & Statistics in the Department of Computer Science. No portion of the work in this thesis has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author.

Signature of Author………

JASMINE KAUR

Table of Contents

Chapter 1

1.1 Introduction

In the realm of statistical modelling, survival analysis methods are applied to model time-to-event data. This dissertation employs the cumulative hazard function as a powerful tool for analysing such data. The dissertation delves deeply into the properties of cumulative hazards, leveraging some of their unique characteristics, including compositional and inverse properties, which have not been previously harnessed for statistical modelling purposes.

These properties serve as the foundation for creating novel univariate families of cumulative hazards and interpretable regression models. Moreover, they facilitate the unification of existing models and families within a cohesive framework. Within this dissertation, a custom distribution is created and introduced, based on the Weibull distribution, using the R package ‘flexsurv’. Additionally, a novel class of mixture models, incorporating frailty and scale parameters, is developed, eliminating the need for explicit integrations.

This research establishes a user-friendly framework for constructing multi-parametric families of cumulative hazards, enabling the modelling of diverse behaviours exhibited by such hazards. It provides intricate insights into the range of behaviours that cumulative hazards can depict and presents a structured framework for combining desired behaviours effectively.

In summary, this dissertation not only introduces innovative modelling approaches but also contributes to a deeper understanding of cumulative hazards and their versatility in capturing various types of behaviours. It offers valuable insights into the analysis of time-to-event data, enhancing the field of statistical modelling and its application to real-world scenarios.

1.2 Aims and Objectives

In the realm of statistical modelling, numerous regression models have been documented within the literature. Despite the availability of various options, many researchers tend to gravitate towards popular models, such as the widely-used Cox Proportional Hazard model. However, this predominant reliance on a limited set of models may not always be the optimal strategy, especially when dealing with diverse datasets or addressing specific research inquiries. Therefore, the principal aim of this dissertation is to undertake a comparative analysis of a broad spectrum of regression models, including those that are less commonly employed.

This research endeavour is motivated by the intention to shed light on the strengths and weaknesses inherent to each of the modelling approaches used. By delving into a wide array of regression models, this research aspires to provide valuable insights that can guide researchers in making more informed choices about which model suits their particular dataset or research objectives. Through this extensive exploration, the research aims to contribute to a deeper understanding of the applicability and suitability of various regression models, thereby enhancing the decision-making process when selecting an appropriate model for specific types of data or research inquiries.

To construct our time-to-event regression models, we intend to utilize the time variable itself along with all other relevant covariates available within the dataset, namely, ‘Age’, ‘Sex’, and ‘Cancer care plan intent’. This comprehensive approach enables us to encompass a wide array of potential predictors that could impact the occurrence of the event of interest. By doing so, we aim to develop more robust and reliable regression models. Ultimately, our objective is to gain a deeper and more nuanced understanding of the factors that exert influence on the timing of events in the context of our study.

Subsequently the primary purposes of this dissertation are outlined below:

Objective 1: Our initial and main objective revolves around a comprehensive comparison regression modelling methodologies, introduced by Rubio et al. (as detailed in the literature review) to our own . To facilitate this comparative analysis, we will leverage the Simulacrum dataset. Our approach to comparing these methodologies will be twofold, aimed at providing a thorough assessment.

Firstly, we will evaluate how well each modelling approach fits the dataset. This assessment entails an examination of the goodness of fit metrics and the statistical properties inherent to each model. We aim to discern which model aligns more closely with the underlying data distribution and captures its nuances effectively. This aspect of our analysis will offer insights into the adequacy of these models in representing the Simulacrum dataset.

Secondly, our comparative analysis will extend to the predictive performance of these models. Specifically, we aim to gauge how effectively each model predicts outcomes when specific variables are omitted from consideration. By undertaking this aspect of the analysis, we aspire to discern the relative strengths and weaknesses of each modelling approach. Finally, we intend to uncover how sensitive these models are to variable exclusion and, consequently, which one exhibits superior predictive capabilities in real-world scenarios.

The overarching goal of this objective is to provide a comprehensive understanding of the merits and limitations of the Rubio et al. regression models in the context of the Simulacrum dataset. Ultimately, we aim to determine which of these approaches is better suited for capturing the intricacies of the dataset and, consequently, for enhancing our insights into the phenomenon under investigation.

Objective 2: Our second objective centres on deriving meaningful conclusions from the Simulacrum dataset, employing the most promising models identified through our comparisons in objective 1. This phase of our research involves a structured approach to model selection and evaluation.

By implementing the selection of procedures and diagnostic assessments, our objective is to ensure that the models chosen in objective 1 are not only the most appropriate but also provide accurate representations of the Simulacrum dataset.

Objective 3: The third goal of this project entails the development of a collection of R functions, leveraging the capabilities of the flexsurv package. These functions are intended to serve as a valuable resource for future cancer research endeavours. Their primary purpose is to empower researchers to conduct time-to-event analyses employing a diverse array of regression models.

Through the development of these functions, our aspiration is to streamline and standardize the process of conducting time-to-event analyses. By doing so, we aim to empower researchers with the means to make more accurate and reliable assessments in their investigations related to cancer and its associated events. These functions will not only contribute to the efficiency of analyses but also have the potential to advance the quality of research outcomes within the domain of cancer research.

1.3 Introduction Of The Data Used

In order to facilitate our comparative analysis, we will make use of the Simulacrum dataset (https://simulacrum.healthdatainsight.org.uk). The Simulacrum is a dataset specifically crafted to simulate patient-like cancer data, serving as a valuable resource for researchers seeking insights into cancer-related studies.

This dataset is meticulously designed to mimic certain data that is securely held by Public Health England's National Cancer Registration and Analysis Service. It's important to note that the data contained within the Simulacrum dataset is entirely artificial and does not contain any information pertaining to real patients. Researchers have the freedom to utilize this dataset for their research, safe in the knowledge that it closely resembles real-world cancer data while posing absolutely no risk of compromising patient confidentiality.

The development of the Simulacrum dataset is the result of a collaborative effort between HDI, AstraZeneca, and IQVIA. This resource was initially made available to the public on November 28, 2018. Since its release, the Simulacrum dataset has proven to be an invaluable tool for researchers in the field of cancer research, offering a secure and ethically sound means of conducting in-depth analysis of cancer data at the record level. Researchers can leverage this dataset to enhance their understanding of cancer-related phenomena without any concerns about patient privacy or data security (HDI, 2023).

Chapter 2

2.1 Basic Introduction To Time To Event Data

In health care research, we frequently quantify the amount of time that passes before an event occurs. For instance, when a patient with asthma is readmitted to the hospital after being discharged, the age at which breastfeeding ceased (Clements et al., 1997), the time from infertility treatment to conception (Luthra et al., 1982), the time to healing of a wound (Nelson et al., 2004), or the time to recurrence of a gallstone (Petroni et al., 2000), are examples of events that may fall under this category. Such data is referred to as time-to-event data. Sometimes the event is unfavourable, like death, and other times it is advantageous, like healing (Bland, 2020).

The most common application of survival data, sometimes referred to as time-to-event data or, more generally, positive-quantity-to-event data, is in survival analysis, where the positive quantity is time and the event is death. For instance, using survival analysis techniques, the time between the diagnosis of a terminal illness and death would be examined in order to identify any potential risk factors and estimate the lifespan of the population. The survival function is given as

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

Where, $S(t)$ is the probability that a person will live past time t,
T is the response variable, $T \geq 0$,
t is the time to event of interest, where t lies within the range from 0 to ∞,
$F(t)$ is the probability that a person would survive for a period of time that is less than or equal to $t,$
Where, F is the Cumulative Distribution Function (CDF).

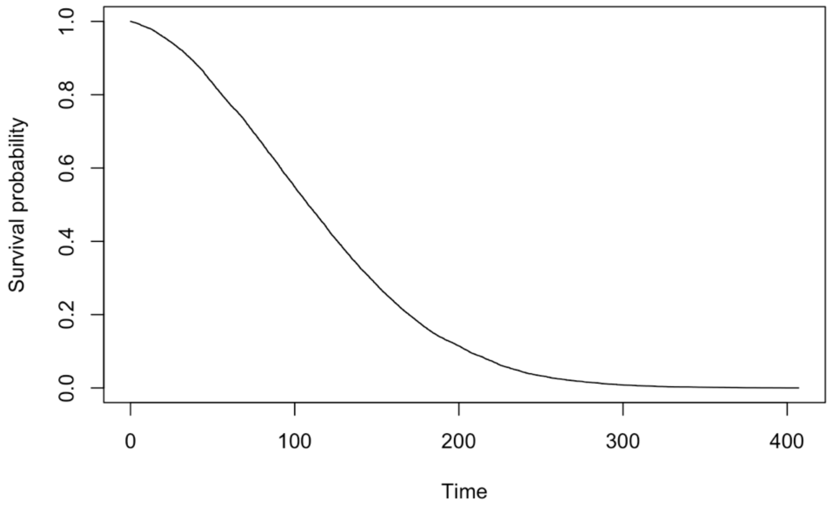

Figure 1: Survival function curve (Soetewey, 2022-12-22)

The curve above displays the percentage of subjects (or experimental units) who, over time, have not encountered an incident. As time passes, events occur, and the number of people who have not witnessed the event drops. (Soetewey, 2022-12-22)

The likelihood of survival of an individual subject to time or a positive quantity t (Moore, 2016), when combined with the hazard function, is typically a primary result of survival analysis.

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

where, $h(t)$ is the Hazard function

This is the likelihood that the event of interest occurs close to time, or a positive quantity t when a subject has survived till t. These two functions are related in the equation below. (McAlpine, 2021)

$$S(t) = \exp\left[- \int_0^t h(u)du\right],$$

2.2 Survival Analysis

The term "survival analysis" refers to the process of analysing data in which the time until the occurrence is of interest. The response is commonly referred to as event time, failure time, or survival time (Columbia University , 2004).

A statistical branch known as survival analysis focuses on estimating survival or, more generally, "times to an event of interest." As previously stated, this event is often interpreted as death, whether it be from a particular condition or natural causes. However, the methods utilized in survival analysis are transferable to many fields outside of biology and medicine, such as engineering, economics, and sociology.

The precise prediction of survival times is one of the primary areas of interest in survival analysis since being able to predict when an event will happen is a very useful skill. Despite the fact that linear models are frequently used in statistical modelling, these methods are no longer appropriate for survival analysis.

One of the essential assumptions in linear modelling is that the data is symmetrically distributed. However, this does not necessarily have to be the case. This assumption is inaccurate since the distribution of survival data is often positive skewed. This sets survival regression apart from linear regression and provides the opportunity for a number of new approaches for developing and analysing models. Statistically, the primary idea is in survival regression is to estimate the survival function S and its relationship with the explanatory variables.

For the purposes of this report, unless otherwise stated, we will consider death to be the event of interest. Typically, this will be a death caused by a major sickness. The continuous time scenario will also be used to build all definitions and models (Thomas, 2016).

2.3 Censoring

When analysing survival data, we often don't have fully knowledge of a person's survival time. As an instance, even if we have some idea of the time of death, we could not have witnessed it. This is referred to as censoring. If we take into account these insights, we can still make use of the incomplete data.

For this research's purpose, we'll primarily concentrate on right censoring. Because of this, the important event occurs after the time that was recorded, and the survival time is really just a lower bound on the real survival time. Right censoring may occur if a patient withdraws from the study or dies from an illness other than the disease of interest.

When we apply right censoring to our data, we make the assumption that the survival time is unaffected by any circumstance that results in data censoring. This isn't always the case, though. (Thomas, 2016)

There are different types of right-censoring:

- **Fixed Type I Censoring**: occurs when a study's intended end point is C years into the follow-up period. In this instance, everyone who does not have an incident observed during the period of the study is censored at C years.
- **Random Type I Censoring**: occurs when the study is intended to finish after C years, however the censoring times for different subjects vary.
- **Type II Censoring**: When a certain number of events have occurred, a study is concluded.

Regardless of the censoring method, we must presume that it is not informative about the occurrence; in other words, the censoring is being done for reasons other than the failure that is about to occur (Columbia University , 2004).

2.4 Survival Analysis Using R

We will utilize R throughout the project to construct survival regression models and carry out the required analyses. We need to use two specialized R packages to achieve this. The 'survival' and 'flexsurv' packages will be primarily utilized (Thomas, 2016). Although, we will be using ‘survival’ package as well, our main focus during the course of this research would be upon the ‘flexsurv’ package as the ‘survival’ package is limited to a small class of models, while ‘flexsurv’ can be used for a vast variety of applications.

‘flexsurv’ is an R program used for fully-parametric survival data modelling. Any parametric time-to-event distribution can be fitted if the user provides a probability density function or hazard function, ideally together with their cumulative versions. The three- and four-parameter generalized gamma and F distributions, among others, are built-in standard survival distributions. Any distribution's parameter can be modelled as a linear or log-linear function of covariates. The program also contains spline model from (P & M, 2002), which allows baseline survival and covariate effects to both be completely adjustable parametric functions of time. The primary model-fitting function, flexsurvreg, makes use of the survreg function from the common survival package and has a familiar syntax (TM, 2016). Censoring and left-truncation are specified in 'Surv' objects. The models are fitted by maximizing the complete log-likelihood, and estimates and confidence intervals for any function of the model parameters can be produced or shown (de Wreede, et al., 2011). (Jackson, 2016).

In addition to these specialized survival analysis tools, we'll use a few more R programs. The charts in this report were made using the 'ggplot2' package, and the 'plyr' function was used to figure out the mean survival times for certain data sets (Thomas, 2016).

In the further part of this section, we are going to discuss the various models that can be analysed and fitted with the ‘flexsurv’ package of R.

Chapter 3

SURVIVAL REGRESSION MODELS

3.1 Cox Model

In simple terms, the Cox proportional-hazards model (Cox, 1972) is a regression model frequently used in medical research to examine the relationship between a patient's survival time and one or more predictor factors (STHDA, n.d.).

A Cox model estimates the treatment effect on survival after controlling for other explanatory variables. Additionally, it enables us to calculate an individual's hazard of dying in light of their prognostic factors. In simple terms, the Cox model is based on a modelling technique for analysing survival statistics. The primary objective of the model is to investigate the effects of multiple variables on survival at the same time. The model enables us to separate the effects of the treatment from the impacts of other variables when it is used to analyse patient survival in a clinical trial. The model could improve the estimation of treatment impact by decreasing the confidence interval. In addition to the time until death, survival times are now frequently used to describe when a specific symptom will appear or when a disease will recur after remission (Walters, 2009).

The Cox model is represented by the hazard function which is denoted by $h(t)$, where the hazard function can be defined as the risk/ hazard of an individual dying at a time t. The hazard function is then given by the following equation:

$$\mathrm{h(t)} = \mathrm{h_0(t)} \times \exp\left(\mathrm{b_1x_1} + \mathrm{b_2x_2} + \cdots + \mathrm{b}_p\mathrm{x}_p\right)$$

3.2 Hazard Function

One of the most crucial ideas in survival analysis is the hazard function h(t), which may be viewed as the instantaneous probability of death at a time t > 0. This function is generated from the probability that a person will pass away at time t if they have been alive up to that point. Let T represent the remaining time before a person's death as a strictly positive random variable. Consequently, the limit depicted below can be used to represent the hazard function: (Thomas, 2016)

$$\mathrm{h}(t) = \lim_{\{\delta \to 0\}} \left\{ \frac{\mathbb{P}(t \leq T \leq t+\delta | T \geq t)}{\delta} \right\}$$

3.3 Cumulative Hazard Function

The cumulative hazard function, which at times is simpler to navigate with than the hazard or survival functions, can be interpreted as the overall risk accumulated up to time t or as the frequency of deaths we would expect to occur throughout a specific time period. This is denoted as

$$\mathrm{H}(t) = \int_0^t h(u)\, du$$

We observe that as $\mathrm{S}(t) \rightarrow 0$ as $\mathrm{t} \rightarrow \infty$, $\mathrm{H}(t) \rightarrow$ as $\mathrm{t} \rightarrow \infty$. Furthermore, because $\mathrm{h(t)}$ is non-negative, $\mathrm{H(t)}$ does not decrease. (Thomas, 2016)

For the research's purpose, we chose to work mostly on cumulative hazard function because it is relatively easier to use and work with.

3.4 Relationships Between Functions

The functions of hazard, survival, and cumulative hazard are all correlated. This means that any unknown function can be obtained if any of the other functions or the probability density function f(t) of the survival times are already known. We can use the following relationships to achieve this:

- $\mathrm{S}(t) = 1 - \mathrm{F(t)}$
- $\mathrm{f}(t) = -S'(\mathrm{t})$
- $\mathrm{h}(t) = \frac{f(t)}{S(t)}$
- $\mathrm{S}(t) = \exp\left(-\mathrm{H}(t)\right))$
- $\mathrm{H}(t) = -\log\left(\mathrm{S}(t)\right))$

$\mathrm{F}(t) = P(\mathrm{T} \leq t)$ demonstrates the cumulative distribution function of T evaluated at t. (Thomas, 2016)

3.5 Common Distributions And Their Hazard Functions

The table below shows the equations for the hazard, survivor, and cumulative hazard functions for particular parametrisations of frequently used distributions in survival analysis. The vector of parameters applied to each given distribution is denoted by 'μ'. Utilizing the relationships mentioned above helps explain the functions in terms of one another in a variety of situations. (Thomas, 2016)

DISTRIBUTIONS	$\boldsymbol{f(t\|\mu)}$	$\boldsymbol{h(t\|\mu)}$	$\boldsymbol{S(t\|\mu)}$	$\boldsymbol{H(t\|\mu)}$
Exponential (λ)	$\lambda e^{-\lambda t}$	λ	$e^{-\lambda t}$	λt
Weibell(λ, γ)bn	$\lambda\gamma t^{\gamma-1} e^{-\lambda t^{\gamma}}$	$\lambda\gamma t^{\gamma-1}$	$e^{-\lambda t^{\gamma}}$	λt^{γ}

Log-logistic(θ, κ)	$\frac{e^{\theta}\kappa t^{\kappa-1}}{(1+e^{\theta}t^{\kappa})^2}$	$\frac{e^{\theta}\kappa t^{\kappa-1}}{1+e^{\theta}t^{\kappa}}$	$\frac{1}{1+e^{\theta}t^{\kappa}}$	$\log\left(1 + e^{\theta}t^{\kappa}\right)$
Log normal(μ, σ)	$\frac{1}{t\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(\log(t)-\mu)^2}{2\sigma^2}\right\}$	$\frac{f(t)}{S(t)}$	$1 - \Phi\ (\frac{(\log\{(t)-\mu)\}}{\sigma}$	$-\log\left(S(t)\right)$
Gamma(θ, λ)	$\frac{\lambda^{\theta}t^{\theta-1}e^{-\lambda t}}{\Gamma(\theta)}$	$\frac{f(t)}{S(t)}$	$1 - \frac{1}{\Gamma(\theta)}\int_0^{\lambda t} u^{\theta-1}e^{-u}d$	$-\log\left(S(t)\right)$
Gompertz(λ, θ)	$\lambda e^{\theta t} exp\left\{\frac{\lambda}{\theta}(1-e^{\theta t})\right\}$	$\lambda e^{\theta t}$	$\exp\frac{\lambda}{\theta}\left(1-e^{\theta t}\right)$	$\frac{\lambda}{\theta}\left(1-e^{\theta t}\right)$

The above table shows the commonly used distributions and functions related to them.

In the above table, $\Gamma(.)$ denotes the Gamma function, where

$$\Gamma(x) = \int_0^{\infty} u^{x-1}e^{-u}du$$

and $\Phi(.)$ denotes the cumulative distribution function for a standard normal distribution.

It's essential to keep in mind that many of these distributions are particular cases of one another. For instance, if $\theta = 1$ the Gamma distribution then is reduced to the Exponential distribution, in addition, if $\gamma = 1$ then the Weibull distribution is again reduced to Exponential distribution. Alongside, when $\theta = 0$, in this case the Gompertz distribution reduces to Exponential distribution. (Thomas, 2016)

3.6 Time To Event Regression

In this section, we're going to describe regression models used in survival analysis. Models used for analysing the relationship between survival time and exploratory variables are called survival regression models. Regression models are used to analyse problems of interest, including comparing survival between groups, assessing/quantifying the impact of covariates, and making predictions about yet-to-be-observed individuals.

Let us assume that the time of survival T is a continuous variable which depends on p number of covariates, $X_1, X_2, \dots, X_p$. For a person, $i = 1, \dots, n$, a set of covariates, we observe $x_i = \left(x_{1i}, \dots, x_{pi}\right)^T$, t_i which is the time of the event/right censoring, and finally, the censoring indicator which is given by

$$\delta_i = \begin{cases} 0, & if\ event\ time\ is\ observed \\ 1, & if\ event\ time\ is\ right - censored \end{cases}$$

This report makes the following assumptions.:

- Event times are always right-censored; no left-truncation is applied. The models can, in theory, be easily generalized to interval-censored and left-truncated data.
- Censorship is both arbitrary and insightful. This shows that the timings of the event and the censoring are distinct events and occur independently , and that the distribution of survival times lacks information on the distribution of censoring times and vice versa (Krstajic, 2017).
- Covariates don't vary over time (Olkin & Martinussen , 2017).

The relationship between the covariates and the distribution of T is described by a survival regression model. Particularly, $H(t|\theta, H_0)$ is the CH function for an individual with covariates x where $\theta = \exp(\beta^T x)$ is the parameter that determines the dependence of the distribution of T on the covariates, β represents a vector of regression coefficients, while CH baseline is represented by H_0, i.e. the CH for an individual who has all covariates equating to zero. Although, $\exp(\beta^T x)$ is often used as the 'link' function, there are other options that might be more appropriate in certain circumstances (Bennett, 1983) (Clayton & Cuzick, 1986).

Significantly, our link function choice guarantees that $\theta > 0$, to ensure that the CH functions used further are valid. We emphasize on fully-parametric models, which assumes a parametric form for the baseline. Furthermore, We commonly write $H(t|\theta)$, as an abbreviation for $H(t|\theta, H_0)$ (Davis, 2018).

3.7 Proportional Hazards Models

According to the proportional hazards (PH) model, covariates and hazard have a multiplicative relationship (Birnbaum & Saunders, 1969) (Hanson & Johnson, 2012)

$$H(t|\theta) = \theta H_0(t)$$

It derives its name from the notion that the ratio of hazard for an individual remain constant over a period of time:

$$\frac{h(t|\theta_1)}{h(t|\theta_2)} = \frac{\theta_1 h_0(t)}{\theta_2 h_0(t)} = \frac{\theta_1}{\theta_2}$$

Given the close correlation between PH models and frailty models (random effects models with a multiplicative impact on the hazard) (Breslow & Day, 1987) θ is often referred to as a frailty parameter. The PH model can be combined with the AFT model, exhibiting a close link between the two (Davis, 2018),

$$H(t|\theta) = H_E^{\theta} \circ H_0(t)$$

3.8 Accelerated Failure Time Models

According to the accelerated failure time (AFT) model, the covariates work in a multiplicatively accelerating/decelerating method to increase/decrease the rate of progression to the event (Birnbaum & Saunders, 1969). The AFT model is represented in terms of the CH function by

$$H(t|\theta) = H_0(\theta t) \tag{1}$$

The parameter θ is known as a scale parameter, since it increases the failure distribution along the time axis.

It is useful to analyse the CH as a composition in the manner of Davis (Davis, 2018) (Bennett, 1983). Compositions will enable us to create 'pairings' between regression models: two regression models form a pair when their CH compositions are reversals. Additionally, writing the CH as a composition can be helpful for data modelling. We have Equation (1) rewritten as

$$H(t|\theta) = H_0 \circ H_E^{\theta}(t)$$

where $H_E^{\theta}(t) = \theta t$ is given as the CH of the $Exp(\theta)$ distribution (Davis, 2018).

3.9 Proportional Odds Models

The proportional odds (PO) model proposes that covariates are multiplicatively related to the event's probability (Aalen, et al., 2008). As a result, the PO model can be expressed in terms of the survival function by

$$\frac{1 - S(t|\theta)}{S(t|\theta)} = \theta\left(\frac{1 - S_0(t)}{S_0(t)}\right)$$

This can be expressed in terms of CH functions as or equivalently using the relation,

$S(t) = \exp(-H(t))$,

$$H(t|\theta) = \log\left(1 + \theta\left(e^{H_0(t)} - 1\right)\right) \tag{2}$$

$$H(t|\theta) = H_{ll} \circ H_E^{\theta} \circ H_G \circ H_0(t)$$

Where $H_G(t) = e^t - 1$ and $H_{ll}(t) = \log(1 + t)$are the CH functions of standard Gompertz distributions and the standard log-logistic respectively, and θ is known as the tilt parameter.

The PO model exhibits intriguing asymptotic behaviour, which can be examined through looking at the asymptotic expansion of Equation (2). For example, as $t \to 0, H(t|\theta) \sim \theta H_0(t)$, the tilt parameter behaves in a similar fashion to the frailty parameter. In contrast, as $t \to \infty, H(t|\theta) \sim H_0(t)$. In the long term, individual hazards converge to the baseline hazard, which is an essential feature of the PO model. This could be a beneficial quality, for instance, the effects of different procedures on various populations are predicted to decrease over time (Aalen, et al., 2008).

The tilt parameter must be $\theta \gg 1$or $\theta \ll 1$, to make a significant impact on the survival/hazard scale (Davis, 2018).

3.10 Kaplan-Meier Estimate Of The Survivor Function

In reality, we might not know which survivor function best fits a particular set of data. Even so, we can still use the data to make inferences by estimating the survivor function. The benefit of this method is that no presumptions on the distribution of survival times are necessary. One such estimate is the Kaplan-Meier estimate.

Let us have n individuals, each having a survival time of $t_1, \dots, t_n$ respectively. Alongside, assuming that there are $r \leq n$ death times along with these survival times. As a result, $n - r$ observations are censored. The death times can be arranged in increasing order to create a series of intervals. These sequential times of death are indicated by $t_{(1)}, t_{(2)}, \dots, t_{(r)}, t_{(1)} < t_{(2)} < \cdots < t_{(r)}$. Additionally, we denote the number of individuals that were living right before $t_{(j)}$ by $n_j, j = 1, \dots, r$ and the number of individuals who die at $t_{(j)}$ by $d_j, j = 1, \dots, r$. We specifically assume that all deaths are unrelated to one another.

The estimated survivor function evaluated at each time $t \in [t_{(k)}, t_{(k+1)})$ merely represents the likelihood of living beyond time $t_{(k)}$. The probability of surviving the interval $[t_k, t_{(k+1)})$ and all preceding intervals is the same as this. Consequently, since the likelihood of surviving in the j^{th} interval can be calculated by

$$\left(\frac{n_j - d_j}{n_j}\right)$$

The following provides the Kaplan-Meier estimate of the survivor function:

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j}\right)$$

Where, $t_{(k)} \leq t_{(k+1)}, k = 1, \dots, r$. Note, $\hat{S}(t) = 1$ for $t < t_{(1)}$, while $t_{(r+1)} = \infty$.

The censored survival times always lies within the intervals $[t_{(j)}, t_{(j+1)}), j = 1, \dots, r$. If the greatest survival time t^* is a censored observation, $\hat{S}(t)$ is undefinable for any $t > t^*$. However, if the longest surviving time is the time of death, then $n_r = d_r$ and $\hat{S}(t) = 0$ for all $t > t_r$. (Thomas, 2016)

We intend to employ the Kaplan-Meier method to visually represent the data. This representation will subsequently be compared against the predictions or fits generated by our parametric models. The Kaplan-Meier method offers a valuable approach for depicting the survival probabilities over time, allowing us to observe the observed survival patterns before juxtaposing them with the anticipated outcomes projected by our parametric models.

3.11 The APGW Distribution For Survival Analysis

The APGW (Adapted Power Generalized Weibull) distribution is a probability distribution that is often used in survival analysis and reliability engineering to model time-to-event data. It is a flexible distribution that can capture a wide range of shapes for survival curves. This distribution, introduced by Burke, Jones, and Noufaily in their seminal work (Burke, et al., 2020), offers a general-purpose framework that can be adapted to various scenarios.

The APGW is a distribution which is defined by four of its parameters. It is defined by its cumulative hazard (CH) function, which is given by:

$$H(t;\phi,\lambda,\gamma,\mathcal{K}) = \lambda\frac{\mathcal{K}+1}{\mathcal{K}}\left[\left(1+\frac{(\phi t)^{\gamma}}{\mathcal{K}+1}\right)^{\mathcal{K}}-1\right]$$

Where, $\phi > 0$, influences the horizontal scaling of the hazard function, and $\lambda > 0$ influences the vertical scaling of the hazard function, while $\gamma > 0,\ \mathcal{K} > -1$ influence the shape of the function.

The APGW distribution's unique feature of having two shape parameters allows it to represent a wide array of hazard shapes in a straightforward manner. These hazard shapes can include constant, increasing, decreasing, U-shaped (up-then-down), and inverted U-shaped (down-then-up). Burke, Jones, and Noufaily argue that the APGW distribution's versatility, which can encompass other well-known distributions like the exponential, Weibull, log-logistic, and Gompertz, coupled with its relative ease of mathematical and computational handling, makes it a superior choice as a flexible and general parametric model when compared to other distributions with two shape parameters, such as the exponentiated Weibull or generalized gamma distributions.

3.11.1 Simulations From The APGW Distribution

The reciprocal or inverse of the cumulative hazard (CH) function of the APGW distribution is expressed as:

$$H^{-1}(\vartheta) = \frac{1}{\phi}\left((\mathcal{K}+1)\left[(1+\frac{\mathcal{K}\vartheta}{\lambda(\mathcal{K}+1)})^{1/\mathcal{K}}-1\right]\right)^{1/\gamma}, \quad \vartheta > 0$$

3.11.2 Simulating Scale Parameter With APGW Baseline

If we intend to simulate from a distribution that is defined by $H(t|\theta, H_0)$, where $H_0 \sim \ APGW(\phi, \lambda, \gamma, \mathcal{K})$, and θ represents a scale parameter, we have

$$H(t|\theta) = \ \lambda \frac{\mathcal{K}+1}{\mathcal{K}}\left[\left(1 + \frac{(\phi\theta t)^{\gamma}}{\mathcal{K}+1}\right)^{\mathcal{K}} - 1\right]$$

Here, this is the CH function of APGW($\phi\theta, \lambda, \gamma, \mathcal{K}$).

3.11.3 Simulating Frailty Parameter With APGW Baseline

When θ is the frailty parameter, then the APGW converts to,

$$H(t|\theta) = \theta\lambda \frac{\mathcal{K}+1}{\mathcal{K}}\left[\left(1 + \frac{(\phi t)^{\gamma}}{\mathcal{K}+1}\right)^{\mathcal{K}} - 1\right]$$

Here, this is the CH function of APGW($\phi, \theta\lambda, \gamma, \mathcal{K}$).

3.11.4 Simulating Tilt Parameter With APGW Baseline

When θ is the tilt parameter, then the APGW converts to,

$$\widetilde{H}_{\theta}(t) = \ H_{ll} \circ \ H_{E}^{\theta} \ \circ \ H_{G}(t) = \log\left(1 + \ \theta(e^{t} - 1)\right)$$

hence, $H(t|\theta) = \ \widetilde{H}_{\theta} \ \circ \ H_0(t)$.

Chapter 4

4.1 Literature Review

In this section, we will comprehensively discuss the extensive body of relevant literature that we have meticulously explored as part of our research process. We will not only elaborate on our systematic approach to searching and selecting pertinent literature but also provide a thorough review and analysis of the chosen works. This literature review encompasses a wide range of sources, ensuring a comprehensive and in-depth examination of the subject matter.

We started with the "Web of Science" website to conduct a systematic literature search for articles on regression analysis of time-to-event data in cancer research. We used primary keywords like "cancer/oncology," "proportional hazard*," "regression," and "survival," as well as secondary keywords like "treatment*," "proportional odds," and "accelerated failure time" to narrow the search. Around 5 million publications were found in the initial broad search using primary keywords, therefore secondary keywords were added to help focus the results. Despite this, relevant papers remained elusive until targeted searches were utilized, which ultimately resulted in the identification of about 13 relevant papers. A careful selection process was then used, eliminating papers that used methodologies unrelated to the focus of my research, concentrating on univariate and fully parametric approaches, and excluding clinical trial data in favour of observational data, multivariate analyses, clinical trial data, and those considered to be too generic or review-oriented. Three relevant publications were ultimately picked, one of which will serve as the main source of analysis while the others will be used as future references. For further guidance, my supervisor also gave me a copy of an unpublished PhD thesis from a former University of Bath student that was pertinent to my research.

The literature review encompasses two primary bodies of literature that we are examining in preparation for our final project. These two approaches are distinguished as follows: the first approach draws inspiration from the regression methodology as presented by Davis, while the second approach is rooted in the regression framework advocated by Rubio et al.

4.2 Classification Of Models By Davis:

Davis in her thesis states that a survival regression model establishes a connection between covariates and the distribution of the event occurrence time T. It does so through a cumulative hazard CH function denoted as $H(t|\theta, H_0)$ or $H(t|X, H_0)$ where $\theta = \exp\left({X_j}^T \beta\right)$ represents a parameter that dictates the influence of covariates on the distribution, β is a vector of regression coefficients, and H_0 is the baseline CH, corresponding to a scenario where all covariates are zero. The choice of $\exp\left({X_j}^T \beta\right)$as the 'link' function is common, but alternative options are available (Bennett, 1983) (Clayton & Cuzick, 1986).

To ensure that CH functions remain valid, we ensure that $\theta > 0$. We mainly focus on fully-parametric models, where we assume a specific parametric form of the baseline. For simplicity, often $H(t|X)$ is used as a shorthand for $H(t|X, H_0)$. Moving forward Davis states the different models stated below:

1. **Proportional Hazards Models (PH):**
   This model asserts that covariates have a multiplicative relationship with the hazard function. The hazard ratio remains consistent over time.

$$H(t|X) = \exp\left(X_j{}^T \beta\right) H_0(t)$$

$$\frac{h(t|X_1)}{h(t|X_2)} = \frac{\left(X_{j1}{}^T \beta_1\right) h_0(t)}{⟦(X⟧_{j2}{}^T \beta_2\big) h_0(t)} = \frac{X_{j1}{}^T \beta_1}{X_{j2}{}^T \beta_2}$$

   Further, the equation $H(t|X) = H_E^{\theta} \circ H_0(t)$ reflects this proportionality.

2. **Accelerated Failure Time Models (AFT):**
   In AFT models, covariates accelerate or decelerate the progression to the event in a multiplicative manner.

$$H(t|X) = H_0\left(\exp\left(X_j{}^T \beta\right) t\right)$$

3. **Proportional Odds Models (PO):**
   PO models relate covariates to event probability multiplicatively. These models are expressed using the survival function given below:

$$\frac{1 - S(t|X)}{S(t|X)} = \exp\left(X_j{}^T \beta\right) \left(\frac{1 - S_0(t)}{S_0(t)}\right)$$

   and have the CH form $H(t|X) = \log\left(1 + \left(\exp ⟦\left(X_j{}^T \beta\right)\right)⟧ \left(e^{H_0(t)} - 1\right)\right)$.

   Here, θ or $\exp\left(X_j{}^T \beta\right)$ is a parameter that can exhibits intriguing asymptotic behaviour.

4. **Proportional Gompertz Model (PGT):**
   The PGT model is derived from the reverse-tilt family within Davis's theoretical framework (Bennett, 1983). It is characterized by a CH function

$$H(t|\theta) = H_0 \circ \left(H_{ll} \circ H_E^{\theta} \circ H_G\right)(t)$$

   Or,

$$H(t|X) = H_0\left(\log\left(1 + \llbracket(\exp\rrbracket\left(X_j{}^T\beta\right))(e^t - 1)\right)\right)$$

This family represents a less common, theoretical approach to regression modelling.

In summary, these different models offer various ways to describe the relationship between covariates and the distribution of event times, each with its own assumptions and characteristics. The choice of model depends on the specific research question and the underlying assumptions about how covariates affect the event of interest over time (Davis, 2018).

4.2.1 Cumulative Hazard Transformations For Generalization Of Models

Davis, in her thesis has also talked about what we refer to as "external time-varying covariates". These are defined as follows:

Let $x(t)$ represent a time-varying covariate, while $X(t) = \{x(s), 0 \leq s < t\}$ represents the covariate history up till time t. Subsequently, we consider $x(s)$ as external if,

$$P\left(s < T \leq s + \varepsilon \middle| T > s, X(s)\right) = P\left(s < T \leq s + \varepsilon \middle| T > s, X(t)\right)$$

for all $0 < s \leq t$ and ε → 0.

The key concept she has highlighted here is that $x(s)$ is linked to the rate at which the event of interest occurs over time, but its future influence up to time $t > s$ is not affected by the event happening at time s. It's worth noting that time-varying covariates may sometimes be deterministic, such as the time of day or predetermined program levels in certain studies. They can also be stochastic processes, like pollution levels in a study on asthma, which affect event times but are not influenced by the occurrence of events.

Given the assumption of smoothness, the she places primary focus on cases where $x(t)$ is a smooth function of time. However, when dealing with stochastic processes, we often observe the covariate only at specific times. To address this, it has been proposed in the literature (Yi-Kuan Tseng, 2005) that the time-varying covariate, $x(t)$ to be represented as below:

$$x_i(t) = \sum_{j=1}^{q} b_{ji}\rho_j(t)$$

as a combination of known basis functions and random effects. Where $\left(\rho_1(t), \dots, \rho_q(t)\right)$ represents the vector of known basis function, while $\left(b_{1i}, \dots, b_{qi}\right)$ denotes the vector of random effects (Davis, 2018).

4.2.3 One-Dimensional Parametric Families

In this section, Davis focuses on constructing one-dimensional parametric families of cumulative hazard functions using fundamental mathematical operations. Specifically, aiming to compose multiple cumulative hazard functions to create new ones. To simplify the expressions, we introduce a notation where if we want to compose H_A^θ followed by H_B^α, it is denoted as

$$H_{A,B}^{\theta,\alpha}(t) := H_A^\theta \circ H_B^\alpha(t),$$

with parameters separated by a comma.

It's important to note that a cumulative hazard function maps a time t to $H(t)$, which is in the range $(0, \infty)$, is still not actually a measure of time but a transformation of the original time scale. We interpret cumulative hazard functions as functional operators to prevent this misunderstanding. Furthermore, we non-dimensionalise the time variable, which permits us to concentrate on the effects of the transformations without modifying the variable's range. For instance, if time is measured in seconds, we divide by one second. This simplifies the visual representation of the functions.

The baseline cumulative hazard is denoted as $H_0(t)$, while the parametric family cumulative hazard, with parameter $\theta = \exp\left({X_j}^T \beta\right)$, is represented as $H(t|X)$. The associated random variable with CH $H(t|X)$ is given as $T_\theta \triangleq \underset{\tau}{\rightarrow} (T_0)$ (Davis, 2018).

1. Families With A Scale Parameter

Further ahead, Davis talks about parametric families, describing the first one as the addition of a scale parameter to the baseline distribution.

Assume that a parametric family's cumulative hazard is

$$H(t|\theta) = H_0(\theta t), \qquad \theta > 0,$$

The parameter θ or $\exp\left({X_j}^T \beta\right)$ is thus referred to as a scale parameter.

This could potentially be expressed as a composition. The family produced by a scale parameter is comparable to

$$H(t|\theta) = H_0 \circ H_E^\theta(t)$$

Where, H_E^θ represents the cumulative hazard associated with an exponential distribution characterized by the rate parameter θ.

2. Families With A Frailty Parameter

Davis talks about another essential parametric family which includes the frailty parameter.

Assuming that a parametric family's cumulative hazard is given by

$$H(t|\theta) = \theta H_0(t), \qquad \theta > 0,$$

The parameter θ is thus referred to as a frailty parameter.

This family is especially significant since it outlines the Proportional Hazards model family, which happens to be the most popular model in survival analysis. These models are frequently referred to as frailty models (Hougaard, 1984), when θ is assumed to be a random variable, and hence, θ is known as the frailty parameter (Marshall & Ingram, 2007).

Introducing a frailty parameter causes the survival function of the fundamental baseline distribution to be elevated by the power of θ, i.e.,

$$S(t|\theta) = S_0(t)^{\theta}$$

Expressing the inclusion of a frailty parameter can be represented as a combination of cumulative hazard functions. The family created through the introduction of a frailty parameter is essentially equal to

$$H(t|\theta) = H_E^{\theta} \circ H_0(t)$$

3. Families Related To A Tilt Parameter

Assuming that a parametric family's cumulative hazard is given by

$$H(t|\theta) = \log\left(1 + \theta\left(e^{H_0(t)} - 1\right)\right), \qquad \theta > 0$$

The θ here is known as the tilt parameter.

This represents a composition in a sequential order that includes the log-logistic, exponential, Gompertz, and baseline cumulative hazard functions, i.e.,

$$H(t|\theta) = H_{ll} \circ \left(\theta H_G\left(H_0(t)\right)\right) = H_{ll} \circ H_E^{\theta} \circ H_G \circ H_0(t)$$

$$= H_{llEG}^{\theta} \circ H_0(t)$$

Where, $H_{ll} \circ H_E^{\theta} \circ H_G(t) = H_{llEG}^{\theta}(t)$ (Davis, 2018)

In conclusion, Davis's exploration of survival regression models and cumulative hazard functions provides a comprehensive framework for understanding and modelling event occurrence over time in the field of survival analysis. Her work sheds light on the intricate relationship between covariates and the distribution of event times, offering researchers a versatile toolkit for addressing various research questions and scenarios.

The models discussed, including Proportional Hazards (PH), Accelerated Failure Time (AFT), Proportional Odds (PO), and the Proportional Gompertz Model (PGT), each present unique approaches to modelling the influence of covariates on event occurrence. These models cater to different research contexts and assumptions, allowing researchers to choose the most suitable one based on the nature of their data and research objectives.

Davis's emphasis on parameterized families, scale parameters, frailty parameters, and tilt parameters enriches our understanding of the flexibility and applicability of survival regression models. By considering these parameters, researchers can fine-tune their models to better capture the dynamics of their data.

Overall, Davis's work serves as a valuable resource for researchers and statisticians engaged in survival analysis. It showcases the importance of thoughtful model selection, parameterization, and understanding the underlying assumptions when studying time-to-event data. As the field of survival analysis continues to evolve, Davis's insights contribute significantly to advancing our knowledge and methodologies in this critical area of research.

Moving forward, we are going to delve into our next literature and talk about the Regression Approach proposed by Rubio et al., providing an examination and analysis of their methodology and findings (Davis, 2018).

4.3 Regression Approach By Rubio Et. Al.:

In this research, Rubio investigated a broad hazard structure that encompasses several specific models, including proportional hazards, accelerated hazards, accelerated failure time structures, and combinations of these models. The study proposes a methodology that utilizes a flexible parametric distribution to implement and apply these diverse hazard models effectively.

To elaborate, the research delves into a comprehensive framework that covers a wide spectrum of hazard models. These models, such as proportional hazards (PH), accelerated hazards (AH), and accelerated failure time (AFT) structures, are commonly used in survival analysis to understand the relationships between covariates and event occurrences over time. What sets this study apart is the development of a method that can seamlessly accommodate and utilize these various hazard models.

To achieve this, Rubio employed a flexible parametric distribution, which serves as the foundation for applying these hazard models. This approach allows us to work with a range of parametric models, making it adaptable to different datasets and research scenarios. By doing so, we enhance the versatility and applicability of these hazard models, enabling researchers to gain deeper insights into time-to-event data and explore the impact of covariates on event probabilities over time.

In the context of this analysis, various excess hazard structures are considered, each dependent on time t and a vector of covariates denoted as X_j. These structures are expressed using the hazard function $h(\quad)$ and the cumulative hazard function $H(\quad)$. To ensure identifiability, Rubio made the assumption that the vector of variables does not include an intercept term. It's important to note that the well-known relationship between survival function $S(t)$ and cumulative hazard $H(t)$ as $S(t) = \exp[-H(t)]$ can be used to derive the survival function based on these hazard structures. The unknown regression parameters are represented by the vector β (Rubio, et al., 2019). Below are the various models discussed in Rubio's work:

1. **Proportional Hazards Model (PH):**
   This model assumes that the hazard function is proportional to a baseline hazard and covariates, with the proportionality determined by regression coefficients β, and is expressed as:

$$H_E^{PH}(t;\ X_j) = H_0(t)\exp\left({X_j}^T\beta\right)$$

2. **Accelerated Hazards Model (AH):**
   In the AH model, the hazard function is modified by both covariates and their exponential form, creating an accelerated or decelerated effect on the hazard. It is expressed as:

$$H_E^{AH}(t;\ X_j) = H_0\left((t)\exp\left({X_j}^T\beta\right)\right)\exp\left(-{X_j}^T\beta\right)$$

   Where, $X_{1j} \subseteq X_j$ and $X_{2j} \subseteq X_j$.

3. **Hybrid Hazards Model (HH):**
   The HH model combines aspects of the PH and AH models. It introduces two sets of covariates, with the first set affecting the baseline hazard and the second set modifying the acceleration effect. It is further expressed as:

$$H_E^{HH}(t;\ X_j) = H_0\left(t\exp\left({X_{1j}}^T\beta_1\right)\right)\exp\left(-{X_{1j}}^T\beta_1 + {X_{2j}}^T\beta_2\right)$$

4. **Accelerated Failure Time Model (AFT):**
   The AFT model directly scales the time variable by the covariates, effectively accelerating or decelerating the event occurrence time, and is expressed as:

$$H_E^{AFT}(t;\ X_j) = H_0\left(t\exp\left({X_j}^T\beta\right)\right)$$

In conclusion, the various models elaborated by Rubio provide different ways to model the excess hazard and its relationship with time and covariates. The choice of a particular model depends on the research context and the underlying assumptions about how covariates influence the hazard function over time. These models allow us to gain insights into the impact of covariates on event occurrence in survival analysis (Rubio, et al., 2019).

Chapter 5

5.1 Description Of The Data Available

The Simulacrum is an accumulation of simulated datasets that replicate the National Disease Registration Service (NDRS) of NHS England. The Simulacrum's datasets are completely constructed from genetic information and imaginary cancer patients' tumor diagnoses as well as treatments. The false records look real, but they don't actually contain any real patient information, making it impossible to identify a specific person using them.

All cancer diagnoses in England are tracked by NDRS, which connects this information to other NHS England data. The Cancer Analysis System (CAS) has a vast and intricate database where all of this data is kept. The information contained herein is extremely valuable and in-depth and pertains to cancer patients' chemotherapy, radiation, and genomic testing regimens. It has the potential to be utilized by academics, pharmaceutical companies, and other researchers to perform research that could improve patient outcomes and provide answers to crucial issues concerning cancer. However, such data also comprises of extremely personal patient information that must be protected, so it is not that easily accessible to the public. Unless special legal and ethical clearances are in place, access to the data can only be made after it has been anonymized.
It was then that the idea of Simulacrum arose. The idea arose was, that if the data held by NDRS was fully anonymised and was made available to anyone publicly, then it would be highly useful for implementing in cancer researches. This would enable study without ever needing to grant people direct data access. people could then construct queries that, after receiving the proper approvals, could be evaluated against the real data and released with complete anonymity.

One of the few important properties of the Simulacrum is that, the Simulacrum accurately preserves many of the statistical characteristics of the original data and has a data structure that is similar to the real data in the CAS. Thus, before submitting a formal request to NDRS to analyse the actual data, it can be used to learn about the structure of the data, create hypotheses, and write code for executing studies.

The Simulacrum v1 dataset that we are employing includes a patient and tumor database as well as the systematic anti-cancer therapy (SACT) dataset. (Simulacrum, 2023)

5.2 Databases And Their Description:

The data set includes a Patient table, a Tumor table, and six different types of SACT datasets, which are as follows:

DATABASE	DESCRIPTION
SIM_AV_PATIENT	This is the main patient table, which contains all the information and details for every patient involved in the data. It contains all kinds of data of an individual patient including information like patient id, sex, ethnicity, etc.
SIM_AV_TUMOUR	This is the main table of the tumour data, and it contains the information and details for each of the tumour registered. It includes information about tumour id, patient id, and a lot more details about the patient's tumour.
SIM_SACT_PATIENT	This table is the SACT table of a patient containing merged patient id and link number which is used for linking between SIM_AV_PATIENT and SIM_SACT_REGIMEN databases.
SIM_SACT_TUMOUR	This table is the SACT table of a tumour which contains information like merged tumor id, merged patient id, consultant specialty code, which is main specialty code of the care professional, primary diagnosis information and more.
SIM_SACT_REGIMEN	This is the regimen table of the SACT which contains merged regimen id which is the drug regimen number, merged tumor id, merged patient id, height at start of regimen, weight at start of regimen, intent of treatment, date decision to treat is the date when the decision to treat the disease was taken, and various other information about the regimen of the tumor.
SIM_SACT_DRUG_DETAIL	This is the SACT table which provides details about the drug used for the treatment, containing information like merged drug detail id, merged outcome id, merged regimen id, merged tumour id, merged patient id, org code of drug provider is the organization code of the drug provider, etc.
SIM_SACT_OUTCOME	This is the SACT table which contains the details about the outcome of the treatment, containing information like merged outcome id, merged regimen id, merged tumour id, date of final treatment, regimen outcome summary which gives the reason for change in the planned treatment, etc.
SIM_SACT_CYCLE	This is the SACT cycle table containing, merged cycle id, cycle number, start date of cycle, opcs procurement code, which is the code of primary procedure, and perf status start of cycle, which gives the performance status of the cycle, and more.

In our research, we rely on two primary databases, namely SIM_AV_PATIENT and SIM_AV_TUMOUR, as the foundational sources for our analysis and investigation. Each of these databases contains essential information that contributes significantly to our study.

SIM_AV_PATIENT serves as the principal patient database, housing critical patient-related data. It encompasses the following key attributes:

❖ PATIENTID: This identifier uniquely represents each patient within the database.

- SEX: gives gender information, categorized into five levels, such as 1 for Male, 2 for Female, and others, offering insights into the patient's sex.
- ETHNICITY: Capturing demographic details, it consists of 71 different levels that characterize patient ethnicity, including categories like White, White British, and more.
- LINKNUMBER: Functioning as an NHS number equivalent, it aids in patient identification.
- DEATHLOCATION: Indicates the location where a patient passed away, featuring 13 distinct levels like Hospital, Private Home, and Hospice Nos.
- NEWVITALSTATUS: Provides the vital status of the patient, encompassing three levels: A for Alive, D for Dead, and X for Exit Posting.
- VITALSTATUSDATE: Records dates of vital events, such as the date of death or the last follow-up for living patients.
- DEATHCAUSE: Comprising various death cause codes, including deathcause_1A, deathcause_1B, and more, each adhering to the ICD-10 format, enabling the classification of causes of death.

Moving on to the second database, SIM_AV_TUMOUR, this repository contains comprehensive data on registered tumors for each patient, acknowledging that a single patient may have multiple tumors. The salient attributes within this database encompass:

- TUMOURID: A unique identifier for each tumor recorded in the database.
- PATIENTID: Correlates with the corresponding patient's ID, establishing a linkage between patients and their tumors.
- DIAGNOSTICDATEBEST: Records the date of the tumor's diagnosis.
- SITE_ICD10: Offers information about the neoplasm's site using 3- and 4-character codes, including SITE_ICD10_02 and SITE_ICD10_02_3CHAR.
- MORPH_ICD10_02: Presents the morphology of the cancer, employing ICD-10-O2 system morphology codes.
- BEHAVIOUR_ICD10_02: Describes the behavior of the tumor, distinguishing between benign, uncertain behavior, and malignant categories.
- STAGE_BEST: Indicates the best 'registry' stage at the time of tumor diagnosis, featuring 37 distinct levels.
- T_BEST, N_BEST, and M_BEST: Flag the best T, N, and M stages as determined by the registry.
- STAGE_BEST_SYSTEM: Specifies the system employed to record the best registry stage at diagnosis.
- GRADE: Reflects the tumor's grade, categorized into 12 levels.
- AGE: Captures the patient's age at the time of diagnosis in years.
- SEX: Reiterates the gender of the patient.
- LINKNUMBER: Serves as a corresponding identifier, matching that of AV_PATIENTS.
- CREG_CODE: Specifies the Cancer registry catchment area code where the patient resided when the tumor was diagnosed.
- SCREENINGSTATUSFULL_CODE: Provides detailed information about the tumor's screening status.
- ER_STATUS, ER_SCORE, PR_STATUS, and PR_SCORE: Offer data on the Estrogen and Progesterone receptor status and scores of the tumor.
- HER2_STATUS: Indicates the HER2 status of the tumor.

- PERFORMANCESTATUS: Records the performance status of the patient at the time of diagnosis.
- CANCERCAREPLANINTENT: Describes the intent of treatment, including categories such as curative (C), non-curative (Z), and not active (X).
- GLEASON: Includes primary, secondary, tertiary patterns of Gleason, along with the combined Gleason primary and secondary scores.
- CNS: Provides information about the Clinical Nurse Specialist.
- ACE27: Contains scores from the Adult Comorbidity Evaluation 27, offering insights into the patient's comorbidity level.
- QUINTILE_2015: Measures deprivation at small area levels, calculated from income domain data in 2015.
- LATERALITY: Specifies the side of the tumor, distinguishing between bilateral, left, midline, and right.
- DATE_FIRST_SURGERY: Records the date of the first surgical event linked to the respective tumor, as documented in the Cancer Registration treatment table.

These two databases, SIM_AV_PATIENT and SIM_AV_TUMOUR, constitute the cornerstone of our research, providing a wealth of patient and tumor-related information essential for our comprehensive analysis and investigation.

Chapter 6

6.1 Exploratory Analysis

In this section, we embark on an exploratory data analysis journey. Our primary objective is to unveil the intricate characteristics of the dataset through a combination of visual representations and statistical measures.

To facilitate our analysis, especially for fully-parametric models, it becomes imperative to establish a parametric form for the baseline. The selection of this baseline parametric form is a crucial decision, given the multitude of possibilities available. The appropriateness of a particular choice often hinges on the specific context of the analysis.

In this chapter, we delve into the analysis and usage of one versatile candidate for the baseline distribution: the adapted power generalized Weibull (APGW) distribution, along with one of the most popular methods for survival regression: Kaplan-Meier. We take this opportunity to implement the APGW distribution, with the aid of the flexsurv package in R.

Our approach involves subjecting the Kaplan-Meier method as well as the APGW distribution to rigorous testing, incorporating a simulation of data from a dataset. Through this comprehensive analysis, we aim to assess the suitability and performance of the APGW distribution along with the Kaplan-Meier analysis and compare it to see which of the both provide a better fit for our models. This exploration not only contributes to the development of robust statistical tools but also provides valuable insights into the behaviour of the APGW distribution in practical applications.

For our research, we specifically focused on utilizing two key databases extracted from the extensive Simulacrum dataset. These databases are identified as SIM_AV_PATIENT and SIM_AV_TUMOUR, each containing a wealth of information relevant to our study.

From these two databases, we carefully selected three specific variables, i.e. Sex, Age, and Cancer care plan intent, that would serve as essential covariates for our exploratory analysis. These covariates form the basis for our analytical approach, which involves employing both the Kaplan-Meier method and APGW regression techniques. These methods will enable us to gain valuable insights into the relationships and patterns within the data.

6.2 Survival Regression Models Using Kaplan-Meier

In this section, we delve into the practical application of several survival regression models, all of which were extensively discussed in Chapter 3 of our dissertation. These models have been carefully selected to address the specific research questions at hand. To perform a comprehensive regression analysis, we implement these models with the Kaplan-Meier method.

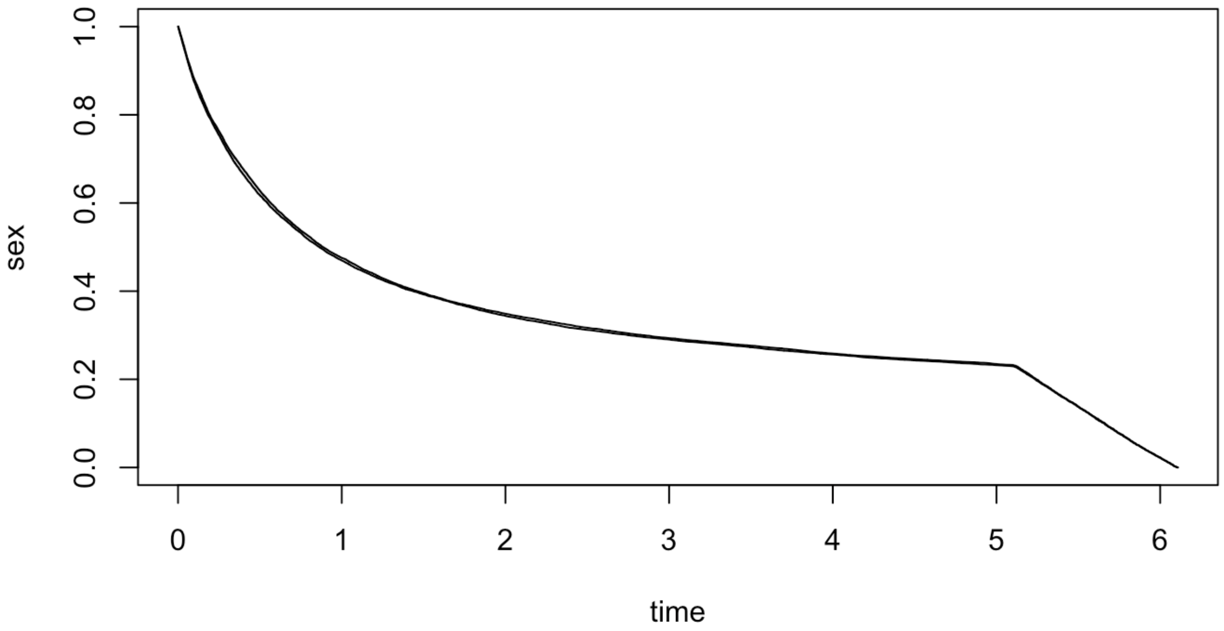

Figure 2: Kaplan-Meier curve for the variable sex

The above depicted Kaplan-Meier curve illustrates the relationship between the variable "sex" and survival over time. This curve presents separate lines for males and females. Upon closer examination, we can discern distinct patterns within the curve.

Firstly, in the initial interval from units 0 to 1, there is a remarkably steep decline in the survival curve. This steep drop suggests a high mortality rate during this early period, indicating a poor survival prognosis.

Secondly, as we progress from units 2 to 5 on the time axis, we observe a more gradual decline in the curve. This rather flatter slope signifies a lower death rate, reflecting a more favorable survival prognosis during this middle phase.

Subsequently, as we move beyond unit 5, the curve takes another sharp decline, indicating another period of elevated mortality. Consequently, the survival prognosis becomes bleak once more.

It is noteworthy that the survival curves for both males and females exhibit strikingly similar patterns. This similarity suggests that both genders share a nearly identical survival experience throughout the observed time frame. This finding underscores the absence of a significant disparity in survival outcomes between the two groups.

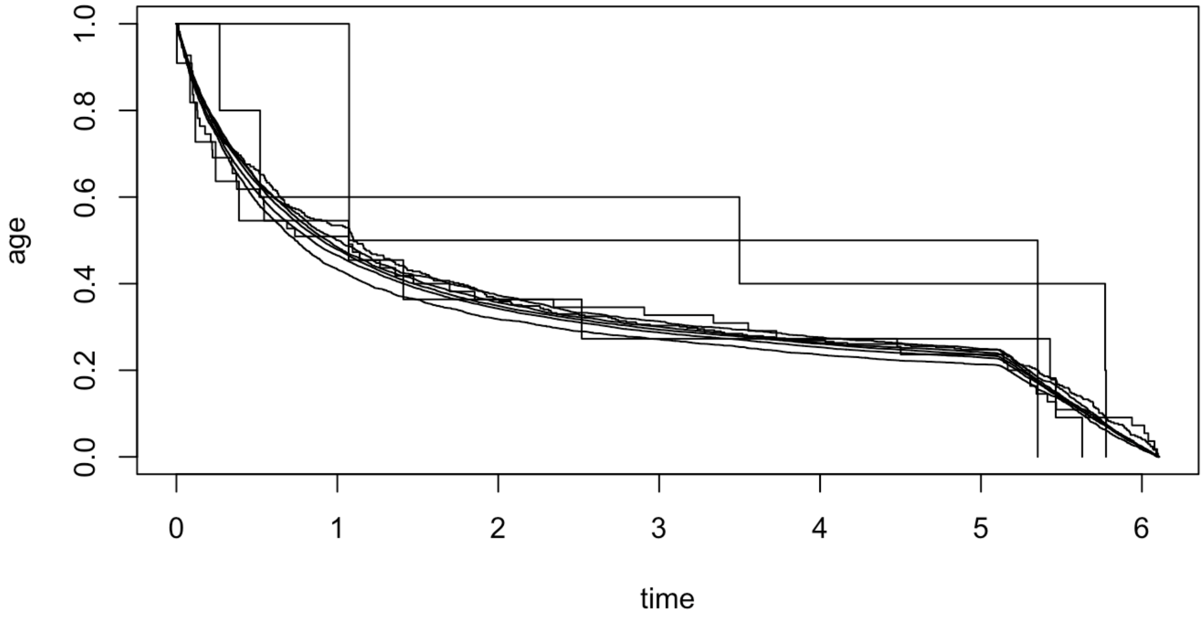

Figure 3: Kaplan-Meier curve for the variable age

The above figure displays Kaplan-Meier curve portrays the relationship between the variable "age" and survival outcomes over time. Notably, the curves represent distinct age groups delineated using the "age_cut" variable. Upon closer examination, we can discern several intriguing patterns within these curves.

Firstly, some of the age group curves exhibit a resemblance to the sex-related survival patterns discussed earlier. Just as with the sex variable, these age groups feature segments that display a steep decline in the survival curve, indicating a high mortality rate. This translates to a rather poor survival prognosis for individuals in these specific age categories.

Conversely, the other age group curves exhibit a more intricate pattern, consisting of both flatter segments and very steep segments. These diverse segments reflect varying death rates over time. The flatter segments suggest a lower death rate, implying a higher likelihood of survival, and thus a more favorable prognosis. Conversely, the steep segments signify periods of pronounced mortality risk, leading to a less optimistic survival prognosis for individuals within those age groups during those time intervals.

What's particularly interesting here is the nuanced nature of these age-related survival curves. Unlike the uniformity observed in the sex-related curves, the age groups display a mixture of survival experiences, with some age groups faring better at certain points in time and others facing higher mortality risk. This underscores the importance of age as a variable in predicting survival outcomes and highlights the need for tailored interventions or strategies for different age groups.

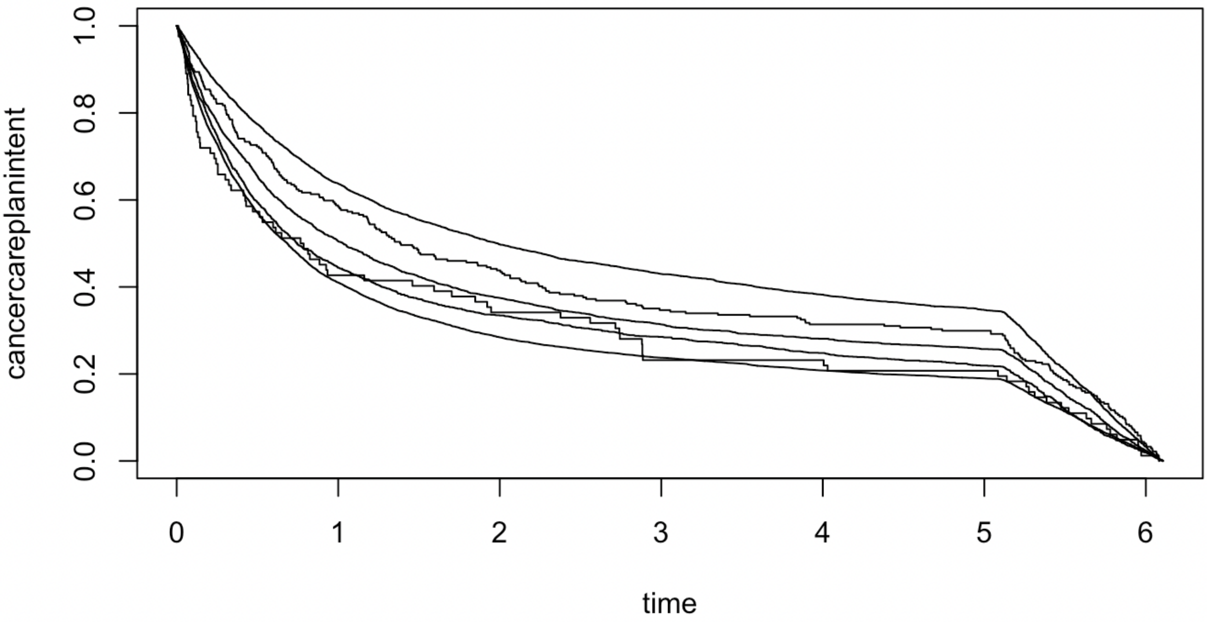

Figure 4: Kaplan-Meier curve for the variable cancercareplanintent

The presented Kaplan-Meier plot illustrates the relationship between the variable "cancercareplanintent" and its impact on survival over time. Each curve within the plot represents different cases or categories within this variable. Upon careful examination, we notice that certain patterns in this plot resemble those observed in the plot of variable sex.

Specifically, we observe recurring patterns that are reminiscent of the trends seen in the sex-related survival analysis. This similarity raises questions about the authenticity of the data being examined. In particular, the consistent nature of these patterns suggest that the data might have been artificially generated or simulated rather than being derived from real-world observations, which confirms that the data we are using is a simulated data rather than a real life one.

6.3 Survival Regression Models Using APGW Baseline

Moving on, we apply the survival regression models explained in Chapter 3, utilizing an APGW baseline method through the flexsurv package. What makes this package particularly useful for our analysis is its capacity to incorporate custom distributions and estimate parametric models while allowing for adaptable covariate effects. To create the custom distribution in R, we utilised Matthew Pawley's code for reference.

APGW Baseline For Sex

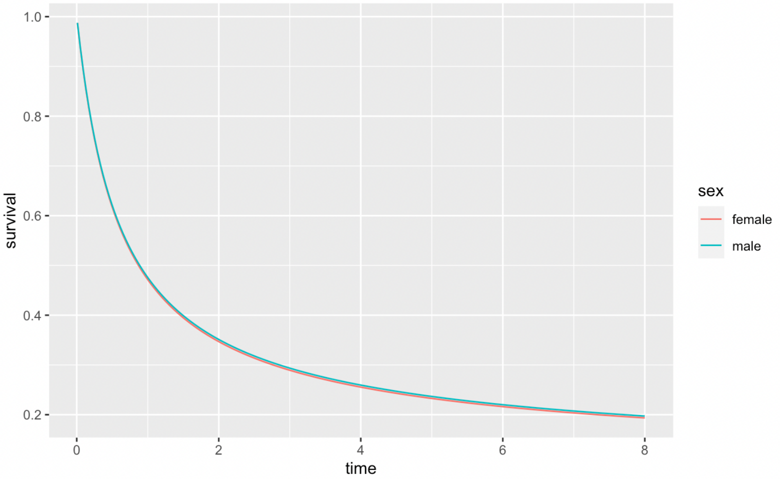

Figure 5: APGW baseline with frailty parameter for variable sex

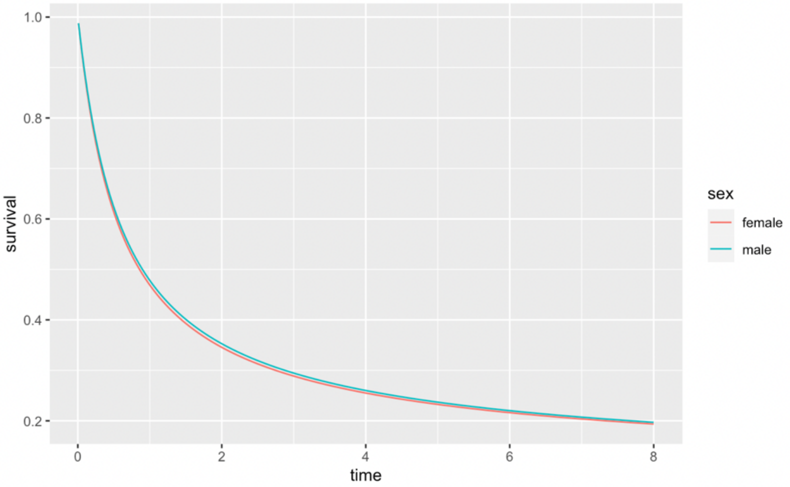

Figure 6: APGW baseline with scale parameter for variable sex

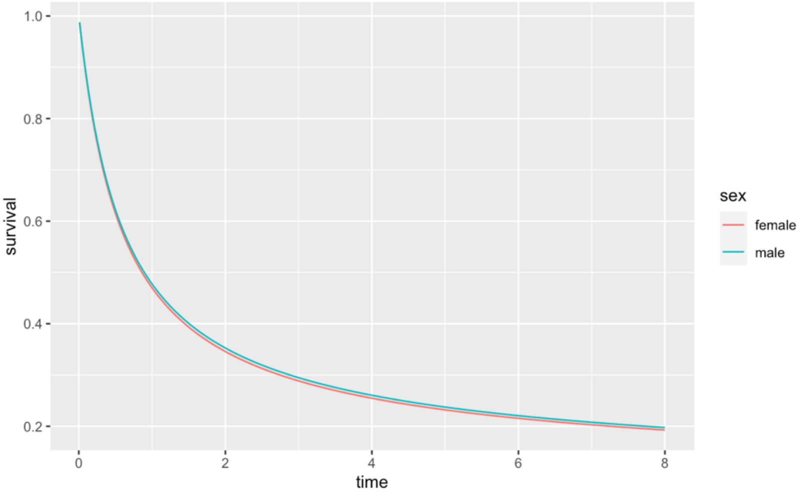

Figure 7: APGW baseline with tilt parameter for variable sex

Figures 5, 6, and 7 present the APGW baseline model for the variable "sex," with parameters of frailty, scale, and tilt, respectively. These plots are color-coded for the male and female curves, and a closer examination unveils intriguing insights into the dynamics of survival analysis.

Initially, a quick glance at all three plots reveals that both male and female curves exhibit a sharp and pronounced decline. This initial steep descent signifies a high mortality rate during the early stages of the observation period, implying a challenging survival outlook during this phase. However, upon closer inspection, a noteworthy transformation becomes evident as we progress along the timeline. The curve's decline becomes more gradual, indicating a decreasing mortality rate. This suggests that over time, both genders experience a reduced risk of mortality, reflecting improved survival prospects as the observation period advances.

What's particularly interesting is the subtle variation in the slopes of all three parameters. While there are slight differences, it's strikingly evident that all three plots share a remarkably similar pattern. This similarity suggests that they exhibit analogous, if not identical, survival characteristics. This coherence across parameters reinforces the notion that, as time progresses, the risk of mortality decreases consistently, regardless of the specific parameter under consideration.

Furthermore, the remarkable similarity between the male and female curves in all three plots is quite striking. This congruence suggests that both genders experience highly similar survival patterns over time, emphasizing the absence of substantial gender-based disparities in survival outcomes within the context of this analysis.

The overall shape of the plots, which exhibits an exponential progression, is a notable feature. This exponential nature suggests that the data follows an exponential distribution, indicating specific characteristics regarding the underlying survival dynamics.

APGW Baseline For Variable Age

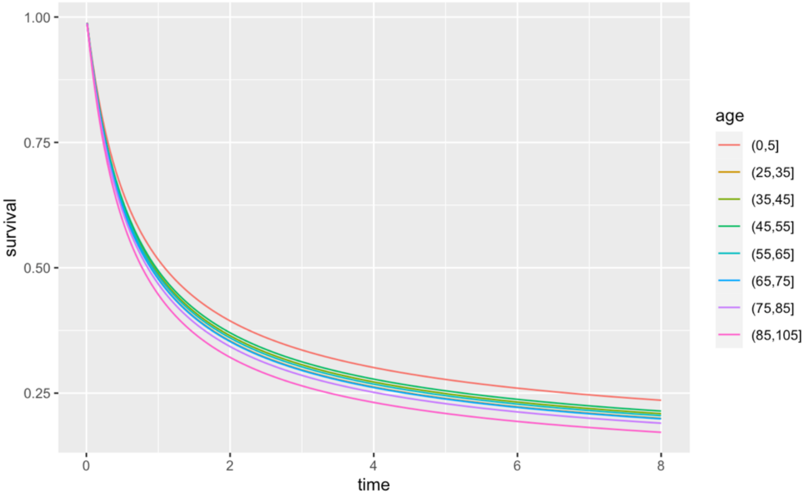

Figure 8: APGW baseline with frailty parameter for variable age

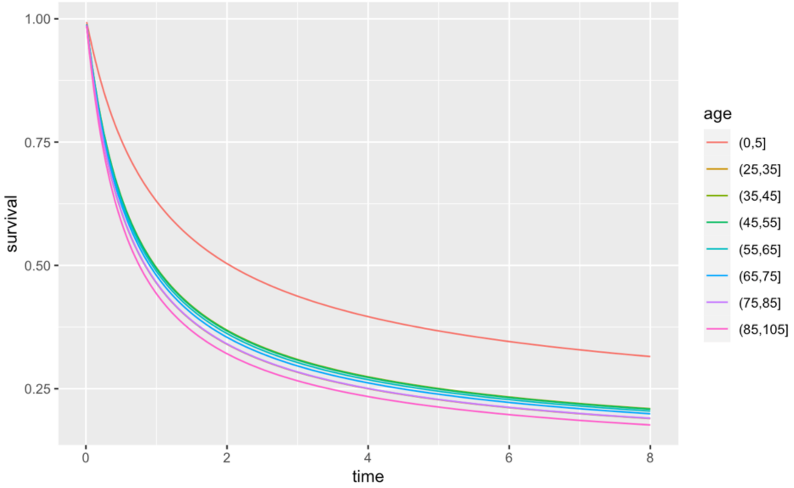

Figure 9: APGW baseline with tilt parameter for variable age

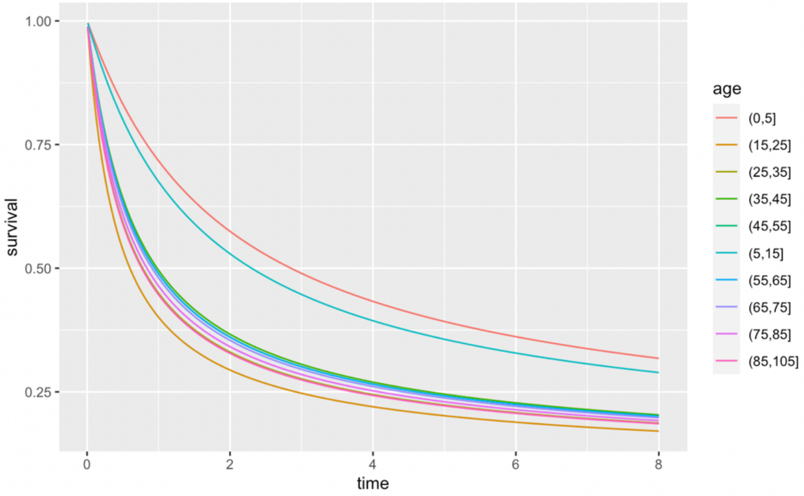

Figure 10: APGW baseline with scale parameter for variable age

Figures 8, 9, and 10 offer insights into the APGW baseline regression model concerning the variable "age." These plots employ various parameters such as frailty, tilt, and scale, respectively, and they showcase survival curves for different age groups, spanning from 0-5 years to 85-105 years.

At first glance, a striking observation across all three plots is the similarity in the nature of survival curves for most age groups. This common pattern suggests that these age groups share analogous survival characteristics.

Firstly, these curves generally exhibit a steep initial slope, gradually smoothing out over time. This trend signifies a high mortality rate at the beginning of the observation period, which subsequently decreases, indicating an improved survival outlook as time progresses.

Another noteworthy observation is that in both plot 8 (frailty parameter) and plot 9 (tilt parameter), the curves follow a clear order based on age groups. The youngest age group (0-5 years) displays the lowest death rate and the highest survival rate, characterized by the shallowest slope. Conversely, the oldest age group (85-105 years) exhibits the highest mortality rate and the lowest survival rate, as indicated by the steepest slope. This ordered progression aligns with our expectations, reflecting the natural aging process and its impact on survival.

However, this age-based trend is notably absent in plot 10 (scale parameter). Here, the curves for different age groups do not follow a specific order. For instance, the last curve represents a relatively younger age group (15-25 years), while the age group of 85-105 years is positioned somewhere in the middle. This lack of a clear order suggests that the scale parameter does not exhibit a consistent relationship with age in terms of survival.

Another key finding is that, across all three plots, the highest survival rates are consistently observed in the youngest age group (0-5 years). This implies that individuals in this age group have the best survival prognosis compared to other age segments.

Lastly, in the age group of 5-15 years, there is an interesting variation between the plots. For frailty (plot 8) and tilt (plot 9) parameters, this age group exhibits a survival pattern similar to the preceding age groups. However, in the case of the scale parameter (plot 10), it displays a survival pattern akin to the youngest age group, suggesting a relatively lower death rate and hence a higher survival prognosis compared to the other age groups.

In summary, these plots provide a comprehensive view of survival dynamics across different age groups and parameter settings. They reveal interesting patterns in mortality rates, the impact of age, and the behavior of different parameters in the APGW baseline regression model.

APGW Baseline For Variable Cancercareplanintent

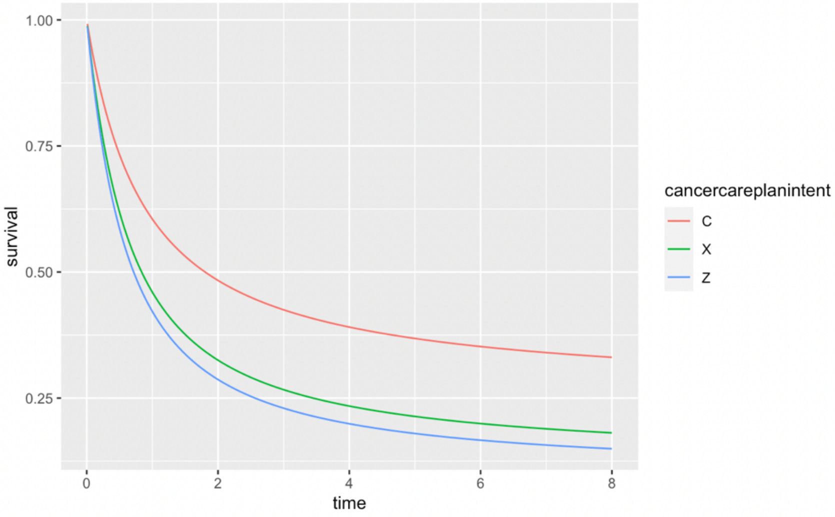

Figure 11: APGW baseline with frailty parameter for variable cancercareplanintent

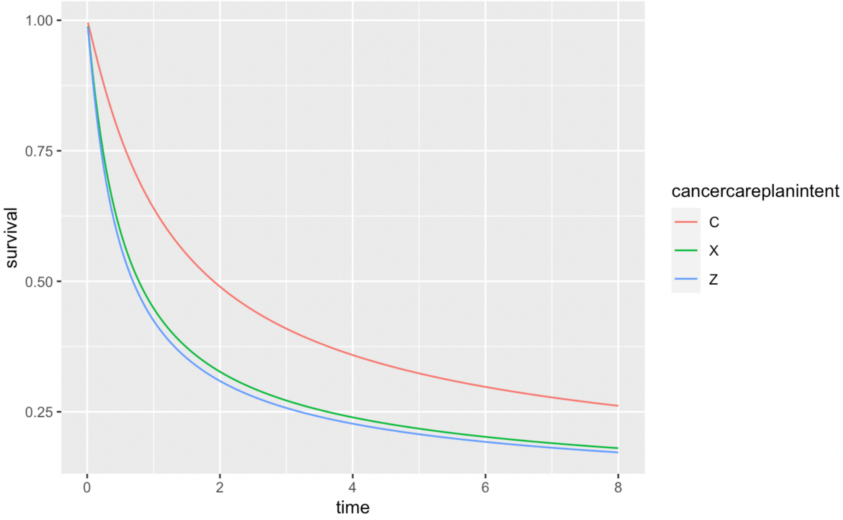

Figure 12: APGW baseline with scale parameter for variable cancercareplanintent

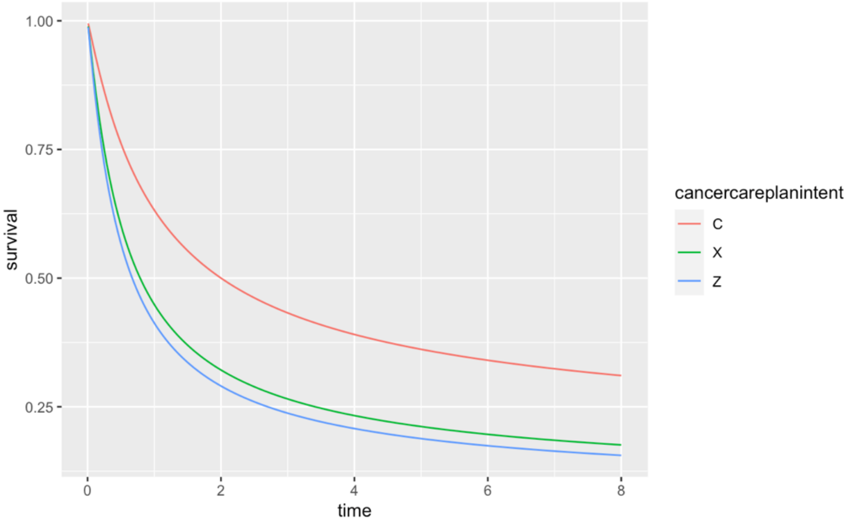

Figure 13: APGW baseline with tilt parameter for variable cancercareplanintent

Figures 11, 12, and 13 delve into the APGW regression model applied to the variable "cancercareplanintent," featuring parameters of frailty, scale, and tilt, respectively. These plots exhibit striking similarities in their overall survival patterns, characterized by an initial sharp decline in survival rates followed by a gradual incline over time.

A prominent and compelling observation is the distinct behavior of the "curative" (C) subgroup, which stands out in all three plots. This subgroup's survival curve exhibits the shallowest slope, indicating a notably low mortality rate and, correspondingly, high survival rates. Notably, this curve becomes progressively smoother as time passes, signifying an increasingly favorable survival outlook. This smooth transition is in line with expectations and suggests that the treatment is effective in enhancing survival.

Conversely, the "non-curative" (Z) subgroup exhibits the opposite trend across all three plots. It consistently displays the highest mortality rate, as evidenced by the steepest slope in the curves. This suggests that individuals in this subgroup face the greatest risk of mortality, and their survival prospects remain challenging over time.

One intriguing observation is related to the "not active or left treatment" (X) subgroup. Surprisingly, this group's survival curve lies between the curves of the "curative" and "non-curative" subgroups in terms of slope and overall behavior. This finding suggests a nuanced survival dynamic for this subgroup, where their survival prospects fall intermediate between those actively receiving curative treatment and those in the non-curative category. This intermediate position might be reflective of various factors affecting their survival outcomes.

Additionally, it's noteworthy that while the survival curves for the "curative" subgroup (C) exhibit similarities across all three parameters (frailty, scale, and tilt), there are variations in the behavior of these curves. Specifically, the curves for frailty and tilt parameters demonstrate greater similarity, while the curve for the scale parameter diverges slightly. This divergence is characterized by steeper slopes in the scale parameter curve, implying a higher death rate compared to the other two parameters. This nuanced difference may hold insights into how these parameters influence the survival dynamics within the "curative" subgroup.

APGW Baseline For Hybrid Hazard

In our forthcoming analysis, we are set to explore APGW regression models that introduce a novel parameter termed the "hybrid hazard." This unique parameter is a fusion of two pre-existing parameters, specifically, frailty and scale. Our approach involves examining individual plots for the APGW baseline model, considering covariates such as sex, age, and cancercareplanintent. Subsequently, we will draw comparisons between these plots and their respective covariates across all other parameters. Additionally, we will juxtapose these findings with Kaplan-Meier plots for the same covariates, offering a comprehensive exploration of the hybrid hazard's impact and its alignment with established survival patterns.

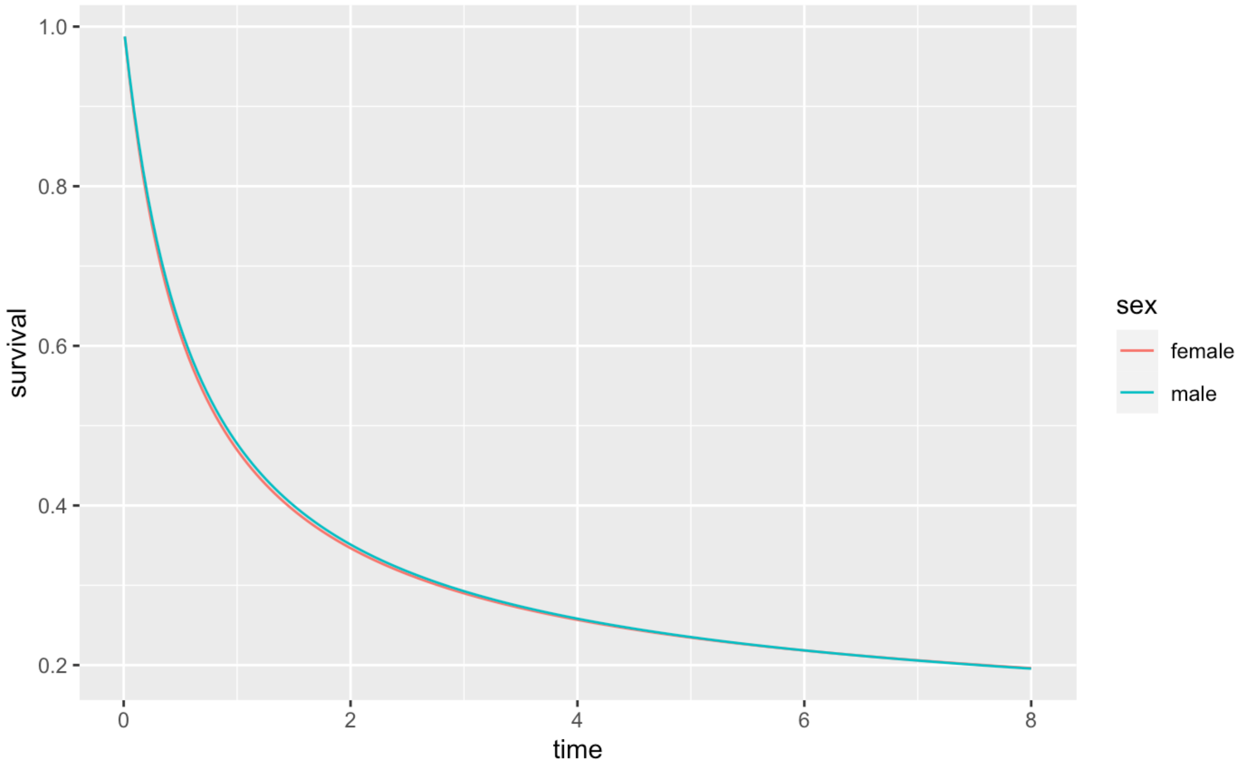

Figure 13: APGW baseline model for hybrid hazard for variable sex

The provided figure illustrates the APGW baseline model employing the hybrid hazard distribution for the variable "sex." Within the plot, distinct colours distinguish the curves representing the two genders. A closer examination of these curves reveals a remarkable resemblance to the patterns observed in the previous plots for parameters like frailty, scale, and tilt.

In essence, the survival curves for this hybrid hazard distribution closely mirror the characteristics we've seen in prior analyses. They exhibit a consistent and familiar nature, showcasing an initial steep decline followed by a gradual incline over time. This observed pattern aligns closely with the survival functions we've previously explored in the context of frailty, scale, and tilt parameters.

The striking similarity between these curves suggests that the hybrid hazard distribution shares common survival dynamics with the previously examined parameters. This consistency underscores the reliability and reproducibility of these survival patterns across different facets of the analysis, particularly highlighting their similarities with the established parameters.

Moreover, when we draw a parallel between this survival function and the Kaplan-Meier plot for the "sex" covariate, we discern some notable distinctions. The Kaplan-Meier curve showcases a distinct pattern with an initial steep decline, a subsequent flattening and smoothing out in the middle, followed by another steep decline towards the end of the observation period. In contrast, the APGW plot reveals a different nature, particularly towards the end of the curve.

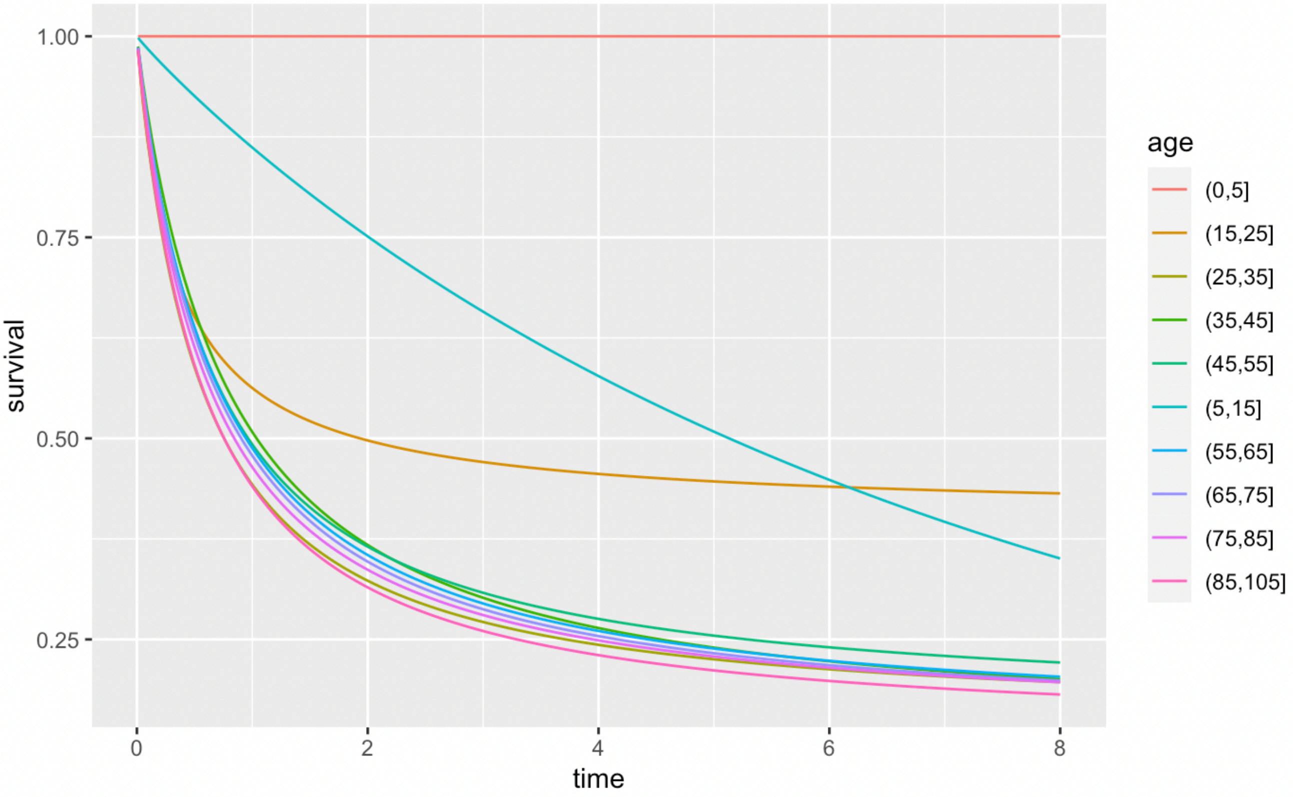

Figure 14: APGW baseline model for hybrid hazard for variable age

The provided figure presents a visual depiction of the APGW baseline model, employing the hybrid hazard distribution and focusing on the "age" variable. Upon conducting a closer examination of these curves, it becomes evident that a significant number of them share striking similarities with patterns observed in previous plots that involved parameters like frailty, scale, and tilt.

Despite the overarching resemblance in the survival patterns across most of the age groups, there are three specific age groups that stand out due to the distinctive nature of their curves. Beginning with the age group 0-5, it's readily apparent that the curve runs virtually parallel to the x-axis. This characteristic signifies an exceptionally low mortality rate and an exceptionally high survival rate. In essence, the flatter the curve, the lower the death rate, resulting in a near 100% survival prognosis.

Turning our attention to the age group 5-15, a notable feature is the steep decline in the curve, reflecting the highest death rate observed among all age groups. This steep descent translates to a relatively poor survival prognosis, making it a distinct outlier.

Lastly, when we examine the age group 25-35, we find that although its curve aligns with the general pattern observed in the majority of curves, it stands out due to its shallower slope compared to the rest. This shallower decline indicates a notably higher survival rate relative to the other similar curves, making it an outlier in terms of a more favorable survival prognosis.

In summary, while most age groups exhibit a consistent survival pattern mirroring that of previous analyses, the age groups 0-5, 5-15, and 25-35 exhibit noteworthy distinctions in their respective survival curves. These deviations in mortality rates and survival probabilities offer valuable insights into the nuanced dynamics of different age groups within the context of the hybrid hazard distribution.

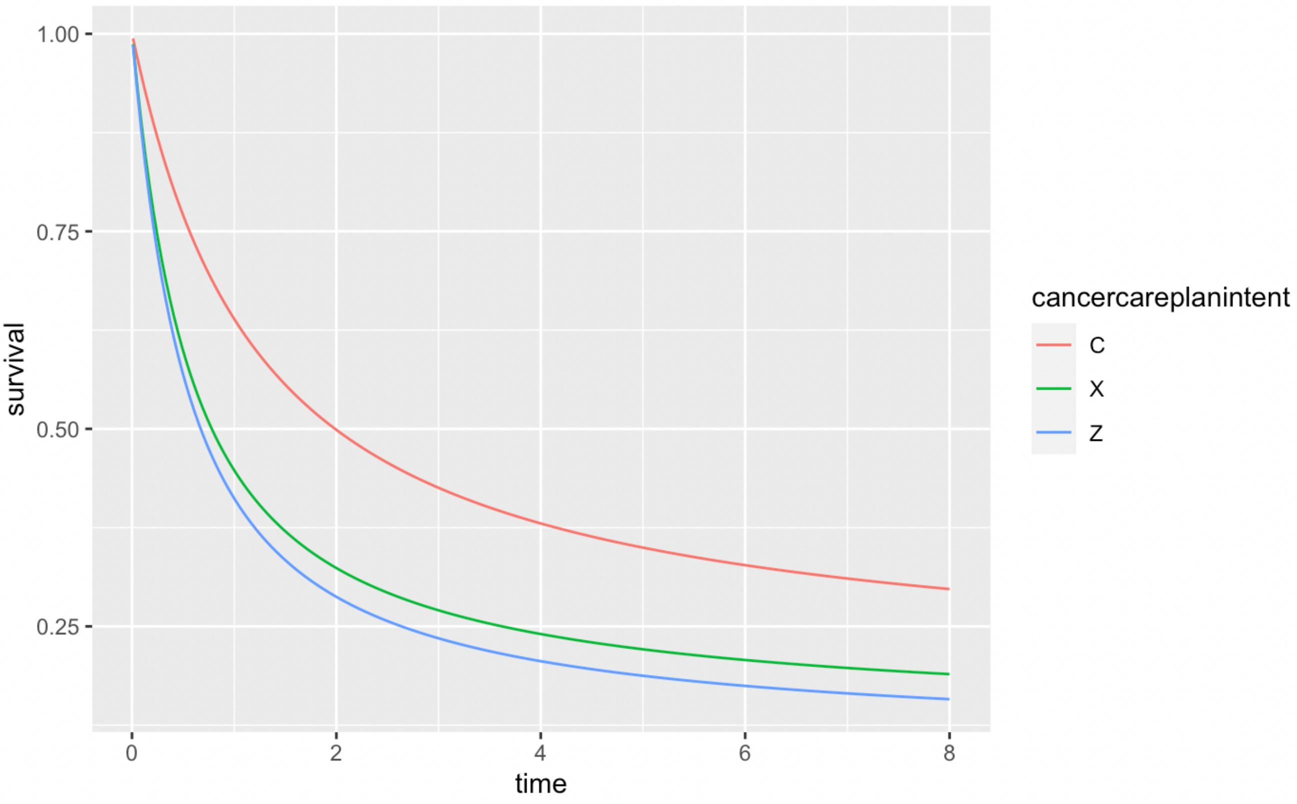

Figure 15: APGW baseline model for hybrid hazard for variable cancercareplanintent

The provided figure offers a visual representation of the APGW baseline model, utilizing the hybrid hazard distribution while focusing on the "cancercareplanintent" variable. The curves in the plot are color-coded to distinguish between various groups within this variable. Upon a closer examination of these curves, we uncover a noteworthy resemblance to patterns observed in previous plots involving parameters such as frailty, scale, and tilt.

At first glance, the survival curves associated with this hybrid hazard distribution bear a strong likeness to what we've observed in our earlier analyses. They follow a consistent and familiar trajectory, characterized by an initial steep descent followed by a gradual recovery over time. However, it's worth noting that this particular plot exhibits a closer similarity with the frailty and tilt parameters compared to the scale parameter.

Additionally, upon juxtaposing it with the Kaplan-Meier analysis of this covariate, we observe that it shows the same properties and differences, akin to what we previously discussed when comparing the covariates of sex and age with the hybrid hazard distribution above.

This differentiation in the survival patterns between the two plots highlights nuances in how they represent the "cancercareplanintent" variable. While both convey insights into survival dynamics, the APGW approach captures certain variations that are not entirely mirrored by the Kaplan-Meier curve. These distinctions may offer valuable insights into the specific characteristics of the hybrid hazard distribution and its impact on survival within the context of this variable.

6.4 Comparison Of APGW Baseline Model Of Hybrid Hazard With The Analysis Of Rubio Et. Al.

Rubio's research offers clarifications on how to interpret parameters estimated from the Accelerated Hazard and Generalized Hazard models, particularly in comparison to more well-known models like the Proportional Hazard and Accelerated Failure Time models. These interpretations are applicable to the Hybrid Hazard model, as it is a specialized case of the GH model. The interpretations depend on the shape of the baseline hazard, which is categorized as either monotonic (increasing or decreasing) or not (bathtub or unimodal).

- For a monotonic baseline hazard:
    - A positive value of a parameter β for a one-unit change in the covariate x indicates that x has a harmful (/beneficial) effect if the baseline hazard is increasing (decreasing).
    - Conversely, a negative value of β means that x has a beneficial (/harmful) effect if the baseline hazard is increasing (/decreasing).
- If the baseline hazard is unimodal (bathtub-shaped):
    - A positive value of β accelerates the hazard's evolution, reaching its maximum (or minimum) sooner.
    - A negative value of β decelerates the hazard's evolution, delaying the attainment of its maximum (or minimum).

Importantly, the AH model maintains the general shape of the hazard but may alter its progression pace. In contrast, the GH model has two parameters, where β_1 modifies the timescale (x-axis) by rescaling time, and β_2 modifies the hazard's magnitude (y-axis), altering its level. The interplay of these parameters can be seen clearly in different hazard shapes.

On comparing the two approaches, we can see that both provide valuable insights into hybrid hazard models, offering distinct perspectives and focusing on different aspects of these models. Each approach contributes uniquely to our overall understanding, and when used together, they can complement each other to provide a more comprehensive view of hybrid hazard models.

The approach and methods used by us in this research is practical and data-centric. It places a strong emphasis on the real-world application of hybrid hazard models to specific covariates, such as "sex," "age," and "cancercareplanintent." This approach provides a visual representation of the hybrid hazard function's behaviour when applied to these covariates. We have discussed the observed survival patterns, highlighting similarities to established parameters like frailty, scale, and tilt

On the other hand Rubio's approach adopts a more theoretical perspective. It delves into the mathematical underpinnings of hybrid hazard models, with a focus on parameter interpretation within the context of AH and GH models. It underscores the significance of parameter signs and their impact on hazard progression. While his approach may appear more abstract, it lays the foundation for a deeper understanding of the inner workings of hybrid hazard models.

When used in conjunction, these two extracts complement each other effectively. The practical insights from our offer a real-world context for understanding how hybrid hazard models can be applied to specific datasets and situations while Rubio's work with its theoretical framework, contributes to a deeper understanding of the mathematical foundations of hybrid hazard models. It offers insights into how model parameters impact hazard behavior, shedding light on the theoretical underpinnings of these models.

Lastly, it's important to acknowledge that the application of the APGW model posed certain computational challenges during the estimation process. Specifically, warnings were encountered during the analysis, indicating that this model demands a more intricate and nuanced approach compared to some of the more conventional models. However, it's crucial to emphasize that these challenges should not overshadow the immense potential and utility that the APGW model brings to the field of survival analysis.

6.5 Conclusion

In conclusion, this dissertation has delved into the fascinating world of survival analysis, with a particular focus on the innovative approach of hybrid hazard models. Through a combination of practical application and theoretical exploration, we have sought to unravel the intricacies of these models and shed light on their utility in understanding survival dynamics within various covariates.

Our journey began by examining Kaplan-Meier curves, which served as our foundation for comprehending the relationship between covariates and survival over time. We carefully dissected the patterns within these curves, uncovering valuable insights into mortality rates, prognosis, and gender-based or age-based disparities in survival outcomes.

The introduction of the APGW baseline model, enriched with parameters such as frailty, scale, and tilt, allowed us to explore survival patterns within covariates like "sex," "age," and "cancer care plan intent." These insights illuminated the dynamic nature of survival, highlighting the evolution of mortality rates over time and underscoring the role of these parameters in shaping survival outcomes. It also brought to out notice that due to all the similarities seen among the different parameters and variables, it clearly shows the signs of the data being simulated rather than a real-life one.

Furthermore, the introduction of the hybrid hazard distribution brought forth an intriguing perspective. This unique parameter, a fusion of frailty and scale, provided us with fresh insights into the dynamics of survival analysis. We observed that the hybrid hazard closely mirrored the characteristics observed in previous analyses, indicating a consistent survival pattern characterized by initial declines and subsequent recoveries over time.

Our comparison with Rubio's research on Accelerated Hazard and Generalized Hazard models offered theoretical depth to our exploration. It emphasized the role of parameter signs and their influence on hazard progression, providing a deeper understanding of the mathematical foundations of hybrid hazard models.

In conclusion, this dissertation has highlighted the versatility and power of hybrid hazard models in analysing survival data. It has showcased its utility in both practical application and theoretical interpretation. While hybrid hazard models may pose computational challenges, their potential benefits for survival analysis cannot be understated.

Link to Github Repository for R code: https://github.com/jasminekaur999/MSc_Dissertation.git

References

STHDA, 2023. *Statistical tools for high-throughput data analysis.* [Online]
Available at: http://www.sthda.com/english/wiki/cox-proportional-hazards-model
[Accessed 02 May 2023].

Thomas, W., 2016. *Parametric Statistical Inference of Survival Regression Models,* Bath: University of Bath.

McAlpine, G., 2021. *A Tutorial on Using Time-To-Event Regression to Analyse Material Strength Data in Mechanical Engineering ,* Bath: University of Bath .

Moore, D., 2016. *Applied Survival Analysis using R,* s.l.: s.n.

Krstajic, D., 2017. How real is the random sensorship model in medical studies?. *Serbia: Research Centre for Cheminformatics.*

Olkin, . I. & Martinussen , T., 2017. Life Distributions. *Springer Series in Statistics.*

Davis, A., 2018. *Modelling Techniques for time-to-event data analysis ,* Bath: University of Bath.

Fleming, T. & Harrington, D., 2005. Counting Processes and Survival Analysis. *Wiley Series in Probability and Statistics .*

Hanson , T. & Johnson, W., 2012. A Bayesian Semiparametric AFT Model for Interval Censored Data. *Journal of Computational and Graphical Statistics ,* Volume 13.2, pp. 341-361.

Hougaard, P., 1995. Frailty Models for Survival Data. *Lifetime Data Analysis 1,* pp. 255-273.

Bennett, S., 1983. Analysis of Survival Data by the Proportional Odds Model. *Statistics in Medicine 2.2,* pp. 273-277.

Simulacrum, 2023. *Simulacrum.* [Online]
Available at: https://simulacrum.healthdatainsight.org.uk
[Accessed 25 April 2023].

Rubio, F., Remontet, L., Jewell, N. P. & Belot, A., 2019. On a General Structure for Hazard Based Regression Models: An Application to Population-based Cancer Research. *Statistical Methods in Medical Research,* 28(8), pp. 2404-2417.

Konstantinidou, D., 2021. *Informative Censoring Models for Liver Transplantation,* Bath: University of Bath .

Soetewey, A., 2022-12-22. *Stats and R.* [Online]
Available at: https://statsandr.com/blog/what-is-survival-analysis/
[Accessed 8 August 2023].

Columbia University , n.d. *COLUMBIA MAILMAN SCHOOL OF PUBLIC HEALTH.* [Online]
Available at: https://www.publichealth.columbia.edu/research/population-health-methods/time-event-data-analysis#:~:text=Time%2Dto%2Devent%20(TTE)%20data%20is%20unique%20because,the%20outco

me%20in%20the%20model.
[Accessed 9 AUGUST 2023].

Jackson, C. H., 2016. flexsurv: A Platform for Parametric Survival Modeling in R. *Journal of Statistical Software,* 70(8), p. 1.

P, R. & M, P., 2002. Flexible Parametric Proportional-Hazards and Proportional- Odds Models for Censored Survival Data, with Application to Prognostic Modelling and Estimation of Treatment Effects. *Statistics in Medicine,* 21(1), p. 2175–2197.

TM, T., 2016. *survival: Survival Analysis.* [Online]
Available at: https://cran.r-project.org/web/packages/survival/index.html
[Accessed 28 July 2023].

de Wreede, L. C., Fiocco, M. & Putter, H., 2011. mstate: An R Package for the Analysis of Competing Risks and Multi-State Models. *Journal of Statistical Software,* 38(7), pp. 1-30.

Cox, D. R., 1972. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological),* 34(2), p. 187–220.

STHDA, n.d. *Statistical tools for high-throughput data analysis.* [Online]
Available at: http://www.sthda.com/english/wiki/cox-proportional-hazards-model
[Accessed 9 August 2023].

Walters, S. J., 2009. *What is a Cox model?.* [Online]
Available at: http://www.bandolier.org.uk/painres/download/whatis/COX_MODEL.pdf
[Accessed 11 August 2023].

HDI, 2023. *Health Data Insight.* [Online]
Available at: https://healthdatainsight.org.uk/about-hdi/
[Accessed 20 August 2023].

Yi-Kuan Tseng, F. H. a. J.-L. W., 2005. Joint Modelling of Accelerated Failure Time and Longitudinal Data. *Biometrika,* 92(3), pp. 587-603.

Birnbaum , Z. W. & Saunders, S. C., 1969. A new family of life distributions. *Journal of Applied Probability,* 6(2), p. 319–327.

Clayton, D. & Cuzick, J., 1986. The semiparametric pareto model for regression analysis of survival times. Papers on Semiparametric Models. *Papers on Semiparametric Models,* p. 19–31.

Breslow, N. E. & Day, N. E., 1987. Statistical Methods in Cancer Research. *International Agency for Research on Cancer.*

Aalen, O., Borgan, O. & Gjessing, H., 2008. Survival and event history analysis : a process point of view. *Springer.*

Hougaard, P., 1984. Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika ,* 71(1), p. 75–83.

Marshall, A. W. & Ingram, O., 2007. *Life Distributions : Structure of Nonparametric, Semiparametric, and Parametric Families.* 1st ed. New York: Springer .

Bland, M., 2020. *University of York.* [Online]
Available at: https://www-users.york.ac.uk/~mb55/yh_stats/surv.htm
[Accessed 2 August 2023].

Columbia University , 2004. *Lecture 15 Introduction to Survival Analysis.* New York: Department of Statistics- Columbia University.

Burke, K., Jones, M. . C. & Noufaily, A., 2020. A Flexible Parametric Modelling Framework for Survival Analysis. *Journal of the Royal Statistical Society Series C: Applied Statistics,* 69(2), p. 429–457.