THE AUSTRALIAN NATIONAL UNIVERSITY

DOCTORAL THESIS

Topics on Survival Analysis with Long-term Survivors

Author:
Muzhi ZHAO

Supervisors:
Prof. Ross MALLER
Dr. Yuguang IPSEN
Prof. Alan WELSH

A thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

for the

Australian National University
Research School of Finance, Actuarial Studies and Statistics

September 8, 2024

Declaration of Authorship

I hereby declare that the work in this thesis is my own except where otherwise stated.

Chapter 6 consists of the paper titled "*Estimation of the Cure Rate for Distributions in the Gumbel Maximum Domain of Attraction Under Insufficient Follow-up*", co-authored with Dr. Mikael Escobar-Bach, Prof. Ross Maller and Prof. Ingrid Van Keilegom. This paper is now published by Biometrika. This chapter also owes very much to the generous discussions with Dr. Yuguang Ipsen. Prof. Ross Maller is credited for the formulation of Theorem 6.5.

The other chapters were completed under the supervision and guidance of Prof. Ross Maller, Dr. Yuguang Ipsen and Prof. Alan Welsh. Dr. Soudabeh Shemehsavar is credited for kindly providing the R codes for the q_n test.

Overall, the author estimates that his contribution to the thesis is no less than 80%.

Signed: Muzhi Zhao

Date: 08 September 2024

Acknowledgements

Throughout my PhD program, I received a great deal of help and support from many people, and I am very grateful to them.

I want to express my sincere gratitude to my supervisor Prof. Ross Maller for his generous and continuing support in many aspects. His encouragements, patience, enthusiasm, and expertise in statistics helped guide me throughout my PhD studies. I met Ross in the last year of my Bachelor degree in actuarial studies. At the time, I was at the crossroads of my life: I wanted to dedicate myself to academia, but I lacked confidence. I was surprised when Ross agreed to supervise my Honours studies, and I am sincerely grateful to him for his encouragement and support that helped me through the downtime of my life. For me, Ross is not only an academic advisor but also a mentor and a friend. He is always caring for people around him, and during my PhD studies, he is always there whenever I need his help. I could not imagine having a better advisor, mentor and friend.

I am also very indebted to my supervisor and friend, Yuguang Ipsen, for her advice, encouragement and support. I am deeply influenced by her enthusiasm for work and her sense of humour. However, the most valuable gift I received from her was the lessons on thinking like a statistician. This benefited me a lot throughout my PhD studies, and I believe it will continue to be valuable for my future career.

I would like to thank my supervisor Prof. Alan Welsh, for many discussions and thoughtful comments on my projects. He is kind and caring. I have also benefited a lot from tutoring his course, Advanced Mathematical Statistics.

During my PhD studies, I was privileged to have met many visitors from overseas. I want to thank Prof. Ingrid Van Keilegom and Dr. Mikael Escobar-Bach, for their advice and discussions that later led to the completion of Chapter 6 of this thesis. I am also thankful to Prof. Sidney Resnick, Dr. Soudabeh Shemehsavar, Prof. David Mason and Prof. Ana Ferreira, for their kind advice and encouragement.

I also want to express my heartfelt gratitude to RSFAS, especially to HDR convener Dr. Tim Higgins and the academic and administrative staff of ANU. I would like to thank Prof. Boris Buchmann and Dr. Yanrong Yang for their advice and encouragement. I am also grateful to my dear friends Dehua Tao, Dr. Adam Nie and many other fellow PhD students in RSFAS for their support.

Many thanks go to my parents for their unconditional love and support throughout my life. None of my achievements could be possible without them.

Special thanks go to my friends Dr. Kevin Jia, Mr. Zack Wei and Mr. Bell Hou for their support and encouragement.

Last but not least, I am so blessed to have Stella as my girlfriend, and no words can express my gratitude towards her. She supported me through the past few years with her continuing help and encouragement. She is always by my side through the highs and downs during my PhD studies. I thank her for being part of my life and her loving care and support in all ways.

Abstract

This thesis focuses on various topics in survival analysis when there are both "immunes" or "cured" individuals, those who will never experience the event of interest, and susceptibles, those who will eventually experience the event of interest, present in the dataset.

Traditional survival models are unsuitable for datasets where immunes are present. In those cases, cure models should be used instead. We introduce some of the commonly used cure models in practice in Chapter 2.

The issues that mixture cure models may face under insufficient follow-up are discussed in Chapter 3. Some existing tests and methods to detect insufficient follow-up are also reviewed.

The generalised F distribution embeds most of the commonly used parametric lifetime distributions as special cases. However, the connections between the generalised F distribution and its sub-models have never been clearly formulated in literature. We fill this gap by clearly outlining such connections in Chapter 4.

In Chapter 5, we provide a detailed methodology for researchers to select a parametric mixture cure model for the data given using a generalised F distribution. Using the results, we demonstrate the application of parametric mixture cured models in practice using real data examples, where we conduct tests for sufficient follow-up, parametric cure model selection and covariate analysis.

Cure proportion estimation is usually of interest to researchers. However, when the assumption that the follow-up period is sufficient does not hold, the parametric cure proportion estimator will be extremely unstable, and the nonparametric cure proportion estimator designed using the Kaplan-Meier estimator will overestimate the cure proportion. In Chapter 6, we develop a nonparametric cure proportion estimator under insufficient follow-up for models in the Gumbel maximum domain of attraction. We show our estimator is asymptotically consistent and normally distributed under reasonable assumptions and performs well in simulation studies with data under both sufficient and insufficient follow-up. We also demonstrate its use with an application to survival data where patients with different stages of breast cancer have varying degrees of follow-up.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

This thesis consists of six main chapters that share a common theme, which is analyzing survival data with long-term survivors using cure models. In this chapter, we present the background knowledge and the underlying motivation for this thesis in Section 1.2. We note that cure models are applied to many different fields of study in practice, and authors in different disciplines may prefer different notations. The notation we use throughout this thesis is introduced in Section 1.3. Section 1.4 reviews some of the preliminary models commonly used in traditional survival analysis. Finally, Section 1.5 provides an outline for the structure of this thesis.

1.2 Background and motivation

In survival analysis, one of the common assumptions is that all of the individuals in the study will eventually experience the event of interest if the given follow-up time is long enough. However, there are instances when a proportion of the population is completely "cured" and therefore is "immune" to the event of interest. In those cases, the cured individuals are referred to as "long-term survivors", and their actual times to event will be infinity. Therefore, the lifetimes of those individuals will necessarily be censored at the limit of follow-up time. We refer to the rest of the population, which will eventually experience the event of interest, as "susceptibles". Without modification, the traditional survival models, such as the Cox proportional hazards model, are not appropriate for survival data with immunes and susceptibles.

Cure models have been developed to analyse such survival data, with the aims of estimating the cured proportion, the overall and susceptible survival functions, the overall and susceptible

hazard functions, and the effects of any covariates on these. The first mixture cure model was introduced by Boag (1949) in a medical context and used to estimate the cured proportion among patients receiving treatment for mouth cancer. It is natural to apply cure models in medical research because, ideally, the patient is indeed cured if a treatment is successful. Although cure models are often applied in a medical context, they are not restricted and are also applicable in other disciplines, such as criminology. For example, Schmidt and Witte (1989) and Broadhurst and Maller (1992) used them to predict criminal recidivism. Moreover, cure models can also be used in economics and social studies. For example, Douglas and Hariharan (1994) studied the hazard of starting smoking for teenagers using the split population model, which is another name for the cure model. Nowadays, there are numerous cure models available, and many authors have expressed their interest in parametric mixture cure models. Peng (1998) introduced the generalised F mixture cure model, which is a flexible model that contains most of the commonly used parametric mixture cure models as special cases. The generalised F distribution has been utilised in model selection, for example, Weng, Yang and Du (2018) selected the optimal model from a range of models contained in the generalised F model using both the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). However, to the best of our knowledge, the relationships between the generalised F model and its sub-models have not yet been clearly structured. Herein, we aim to clearly outline such relationships and apply them in mixture cure model selection.

In an experimental situation with censored survival times, it may be hard to distinguish long-term survivors from susceptibles, and the challenge is to make inferences about the proportion of long term survivors in the population from a sample of possibly censored survival times. In doing this, it is important to take into account features of the analysis which without attention may lead to misleading or meaningless results on various aspects of interest, for example, the estimates of the cured proportion. Among these features, a common assumption, often made implicitly in analysis, is that the follow-up in the sample has been sufficient to allow long term survivors in a trial to become apparent. Maller and Zhou (1996) address this issue. Their criterion is to decide sufficient follow-up when the censoring time distribution has its right extreme greater than or equal to the right extreme of the distribution of the susceptibles' survival times. When a parametric model is fitted and there is insufficient follow-up, it will be hard to distinguish between a high cure proportion with a short tail for the susceptible survival function and a low cure proportion with a long tail, as noted by Li, Taylor and Sy (2001). Motivated

by this problem, the other objective of this thesis is to develop methods to analyse survival data with cure proportion under insufficient follow-up.

1.3 Notations

Typical survival data can also be referred to as lifetime data, as noted by Lawless (1982). Collett (2003) defined survival data as measurement of the length between a well-defined time origin and the time for the occurrence of some particular event of interest for each individual.

Throughout this thesis, if not otherwise specified, we assume a random censorship model (Gill 1980, p.21) with right censoring. We consider a sample of size n on a population consisting of two types of individuals – susceptibles and immunes, in proportions p and $1-p$, $0 < p \leq 1$. Only susceptibles can experience the event of interest. Their lifetimes (survival times) have distribution F_0 on $[0, \infty)$, a proper distribution, but may be censored according to the distribution G on $[0, \infty)$, also a proper distribution. Immune or cured individuals, on the other hand, live forever, but their lifetimes in the sample are always censored, also according to the distribution G.

To model this situation we associate with the ith individual a latent Bernoulli random variable B_i such that $P(B_i = 1) = p = 1 - P(B_i = 0)$ and positive random variables T_i^* and U_i. Individuals with $B_i = 1$ are susceptible and have survival times distributed as F_0. Individuals with $B_i = 0$ are immune, and have T_i^*, formally, equal to ∞. Survival times of both types are subject to censoring by random variables U_i with distribution G. Whether or not an individual's survival time is censored is recorded, but whether individuals are susceptible or cured is not known to the experimenter. Thus the data in a sample of size n consists of observations on the sequence of 2-vectors $(T_i = \min(T_i^*, U_i),\ C_i = I\{T_i^* \leq U_i\}; 1 \leq i \leq n)$ (with $I\{\cdot\}$ denoting the indicator function). The random variables T_i^* and U_i are assumed independent of each other, for each i, and also from T_j^* and U_j, $1 \leq j \leq n$, $j \neq i$. In addition to this, when covariates are not introduced (for example, in Chapter 6), we assume that the T_i^* all have the same F_0 and the U_i all have the same G. In this case, T_i and U_i are independent and identically distributed.

In Chapters 2 and 5, we are also interested in the effect of covariates $\mathbf{X}$ on the overall survival model and cure proportion estimations. In these cases, we let the distribution of each T_i depend on the covariate vector using logistic regressions (with the vector of coefficients $\mathbf{b}$) and appropriate link functions within the generalised linear model framework (with the vector of

coefficients $\boldsymbol{\beta}$). Overall, the observed survival data for the i-th individual would be

$$\mathcal{D}_i := \{T_i, C_i, \mathbf{X}_i\}.$$

As aforementioned, the aims of cure models include estimating the cure proportion $1 - p$, the overall survival model with a cumulative distribution function $F(t)$, the susceptible survival model with a cumulative distribution function $F(t|B = 1) = F_0(t)$. For this thesis, we denote the survival functions, which are the tail of the cumulative distribution functions, as $S = 1 - F$ and $S_0 = 1 - F_0$. We are also interested in the overall hazard rate $h(t)$ and cumulative hazard $H(t)$, along with the susceptible hazard rate $h(t|B = 1) = h_0(t)$ and cumulative hazard $H(t|B = 1) = H_0(t)$.

Most of the cure models are constructed based on traditional survival models. With the notations defined, it is important to review some standard survival models that are widely used in practice before introducing cure models. We note that traditional survival analysis assumes $p = 1$, and hence F_0 is equivalent to F.

1.4 Review of traditional survival models

1.4.1 Nonparametric survival models

The Kaplan-Meier estimator is the most well-known nonparametric estimator of the survival distribution function from incomplete observations. When there are no covariates, we have only available the possibly censored survival times T_i of n patients and their corresponding censoring indicator C_i. The estimator is defined by Kaplan and Meier (1958) as follows.

Theorem 1.1 (Kaplan-Meier Estimator)**.** *Let $\tau_{(1)} < \tau_{(2)} < ... < \tau_{(k)}$ denote the distinct ordered uncensored survival times, where k is the total number of uncensored observations. The estimator of the probability of survival for each interval $[0, \tau_{(1)})$, $[\tau_{(1)}, \tau_{(2)})$,..., $[\tau_{(k-1)}, \tau_{(k)})$, $[\tau_{(k)}, \infty)$ is the ratio between the total number of deaths during each interval and the total number of individuals at risk. Hence, the Kaplan-Meier estimator of the survival function is*

$$\hat{S}_{KM}(t) = \prod_{j:\tau_{(j)}<t} \left(1 - \frac{\sum_{i=1}^{n} \mathbf{1}(T_i = \tau_{(j)}, C_i = 1)}{\sum_{i=1}^{n} \mathbf{1}(T_i \geq \tau_{(j)})}\right),$$

where $t > 0$ and $\mathbf{1}(A)$ is the indicator function with $\mathbf{1}(A) = 1$ if the event A is true and $\mathbf{1}(A) = 0$ otherwise. Intuitively, it is the same as

$$\hat{S}_{KM}(t) = \prod_{j:\tau_{(j)}<t} \left(1 - \frac{\textit{Number of Failures at Time } \tau_{(j)}}{\textit{Total Number at Risk at Time } \tau_{(j)}}\right).$$

The corresponding estimator of the distribution function F will be denoted as

$$\hat{F}_n(t) = 1 - \hat{S}_{KM}(t).$$

The variance of $\hat{S}(t)$ can be estimated using Greenwood's method

$$\widehat{Var}\big(\hat{S}_{KM}(t)\big) = \hat{S}_{KM}(t)^2 \sum_{j:\tau_{(j)}<t} \frac{\sum_{i=1}^{n} \mathbf{1}(T_i = \tau_{(j)}, C_i = 1)}{\sum_{i=1}^{n} \mathbf{1}(T_i \geq \tau_{(j)}) \left(\sum_{i=1}^{n} \mathbf{1}(T_i \geq \tau_{(j)}) - \mathbf{1}(T_i = \tau_{(j)}, C_i = 1)\right)}.$$

The Kaplan-Meier product-limit estimator can be viewed as an analogue of the empirical distribution function (EDF) for censored data to estimate the distribution of the survival times.

1.4.2 Parametric survival models

There are numerous distributions that are naturally suitable for modelling lifetime data. Examples include the generalised gamma distribution, the exponential distribution, the gamma distribution, the Weibull distribution, the log-normal distribution and the Rayleigh distribution. Among them, the generalised gamma distribution is a three-parameter gamma distribution that contains as sub-models those distributions and many others. It was introduced by Stacy (1962), and its density function, with parameters $\underset{\sim}{\theta} = \{\alpha, \lambda, \gamma\}$, can be expressed as

$$f_{gg}(t) = \frac{\alpha}{\Gamma(\gamma)} t^{\alpha\gamma-1} \lambda^{\alpha\gamma} \exp\big(-(\lambda t)^{\alpha}\big). \tag{1.1}$$

The corresponding survival function for the generalised gamma distribution can be expressed in terms of the incomplete gamma functions. We first define the lower incomplete gamma function as

$$I(\gamma, x) = \frac{1}{\Gamma(\gamma)} \int_0^x u^{\gamma-1} \exp(-u) \mathrm{d}u.$$

Applying the transformation $z = (\lambda u)^\alpha$ to (1.1) results in

$$\begin{aligned} F_{gg}(t) &= \int_0^t \frac{\alpha\lambda}{\Gamma(\gamma)} (\lambda u)^{\alpha\gamma-1} \exp\big(-(\lambda u)^\alpha\big) \mathrm{d}u \\ &= \int_0^{(\lambda t)^\alpha} \frac{1}{\Gamma(\gamma)} z^{\gamma-1} \exp(-z) \mathrm{d}z \\ &= I\big(\gamma, (\lambda t)^\alpha\big). \end{aligned} \tag{1.2}$$

The generalised gamma distribution is a flexible family of distributions, and it includes most of the commonly used lifetime parametric distributions as special cases. The relationships between the generalised gamma distribution and its sub-models are rather obvious, except for the log-normal distribution, which we will show later in this section.

List of Lifetime Distributions					
Lifetime Distributions	α	λ	γ	$S(t)$	$f(t)$
Weibull	α	λ	1	$\exp\{-(\lambda t)^\alpha\}$	$\alpha\lambda(\lambda t)^{\alpha-1}\exp\{-(\lambda t)^\alpha\}$
Exponential	1	λ	1	$\exp(-\lambda t)$	$\lambda \exp\{-(\lambda t)\}$
Log-Normal	α	λ	$\gamma \to \infty$	$1 - \Phi\Big[\frac{\log t - \big((\log\gamma)/\alpha - \log\lambda\big)}{(\alpha\sqrt{\gamma})^{-1}}\Big]$	$\frac{\alpha\sqrt{\gamma}}{(2\pi)^{1/2}t} \exp\Big[\frac{\big(\log t - \big((\log\gamma)/\alpha - \log\lambda\big)\big)^2}{2(\alpha\sqrt{\gamma})^{-2}}\Big]$
Gamma	1	λ	γ	$1 - I(\gamma, \lambda t)$	$f(t) = \frac{t^{\gamma-1}\lambda^\gamma}{\Gamma(\gamma)} \exp\big(-(\lambda t)\big)$
Rayleigh	2	λ	1	$\exp(-\lambda^2 t^2)$	$2\lambda^2 t \exp\{-(\lambda t)^2\}$

TABLE 1.1: Sub-Models of the generalised Gamma Model

In practice, the generalised gamma distribution may suffer from identifiability issues since different parameterisations may result in like shaped densities. In addition, the asymptotic normality of the maximum likelihood estimators can be arrived at rather slowly, even for sample sizes as large as 400, as noted by Prentice (1974) and Lawless (1982).

There are also a number of parametric lifetime distributions that are not sub-models of the generalised gamma distribution; they include the log-logistic distribution, the Gompertz distribution, the inverse Gaussian distribution, the truncated normal distribution (on the left of 0). The most commonly used model among them is the log-logistic distribution. The log-logistic distribution and the generalised gamma distribution are sub-models of the generalised F distribution, as noted by Peng, Dear and Denham (1998). Further discussion on the generalised F distribution is presented in Chapter 4.

It was noted by researchers that the generalised gamma distribution may often encounter

convergence problems while using methods such as the Newton-Raphson algorithm to find the maximum likelihood estimators, even with a moderate sample size of 200. To solve this problem and show that the generalised gamma distribution contains the log-normal distribution as a limiting case, Prentice (1974) proposed the log-gamma model and extended generalised gamma distribution.

Let us consider the transformation $W = \alpha \log(\lambda T)$ for T following the generalised gamma distribution. Then

$$\begin{aligned} f_W(w) &= f\big(\exp(w/\alpha)/\lambda\big) \times \Big|\frac{dT}{dW}\Big| \\ &= \frac{\lambda\alpha}{\Gamma(\gamma)}(\exp(w/\alpha))^{\gamma\alpha-1}\exp\Big[-(\exp(w/\alpha))^{\alpha}\Big] \times \frac{\exp(w/\alpha)}{\lambda\alpha} \\ &= \frac{1}{\Gamma(\gamma)}\exp(w/\alpha)^{\gamma\alpha}\exp\Big[-(\exp(w/\alpha))^{\alpha}\Big] \\ &= \frac{1}{\Gamma(\gamma)}\exp\big(\gamma w - \exp(w)\big). \end{aligned}$$

This one-parameter gamma distribution is called the "log-gamma distribution" by Bartlett and Kendall (1946). Note that the relationship between a log-gamma distribution and a gamma distribution is different from that between a log-normal distribution and a normal distribution. The random variable W has moment generating function

$$M_W(\theta) = \mathbb{E}[e^{\theta w}] = \frac{\Gamma(\theta+\gamma)}{\Gamma(\gamma)}.$$

Hence, the expected value and variance of W are

$$\begin{aligned} \mathbb{E}[W] &= \frac{d\log\Gamma(\gamma)}{d\gamma} = \psi(\gamma), \\ Var[W] &= \frac{d^2\log\Gamma(\gamma)}{d\gamma^2} = \psi'(\gamma), \end{aligned}$$

where the digamma function $\psi(\gamma)$ and the trigamma function $\psi'(\gamma)$ can be expressed in the form

$$\begin{aligned} \psi(\gamma) &= \log\gamma - \frac{1}{2\gamma} - \frac{1}{12\gamma^2} + \frac{1}{120\gamma^4} - \frac{1}{252\gamma^6} + \cdots \\ \psi'(\gamma) &= \frac{1}{\gamma} + \frac{1}{2\gamma^2} + \frac{1}{6\gamma^3} - \frac{1}{30\gamma^5} + \cdots \end{aligned}$$

For large γ, i.e., $\gamma \to \infty$, the first term of each series dominates.

This suggests considering the transformation $Z = (W - \log\gamma)\sqrt{\gamma}$. The probability density function of Z is

$$\begin{aligned} f_Z(z;\gamma) &= f_W(Z/\sqrt{\gamma} + \log\gamma) \times \left|\frac{dW}{dZ}\right| \\ &= \frac{1}{\Gamma(\gamma)} \exp\left(\gamma(z/\sqrt{\gamma} + \log\gamma) - \exp(z/\sqrt{\gamma} + \log\gamma)\right) \times \frac{1}{\sqrt{\gamma}} \\ &= \frac{\gamma^{\gamma-1/2}}{\Gamma(\gamma)} \exp\left(\sqrt{\gamma}z - \gamma\exp(z/\sqrt{\gamma})\right). \end{aligned} \tag{1.3}$$

As $\gamma \to \infty$, the distribution of the re-scaled variate $Z = (W - \log\gamma)\sqrt{\gamma}$ converges to the standard normal distribution. To see this relationship, Lawless (1982) suggested first expanding $\exp(z/\sqrt{\gamma})$ in a power series, so that

$$\exp(z/\sqrt{\gamma}) = \sum_{n=0}^{\infty} \frac{z^n}{n!\gamma^{n/2}}.$$

After rearranging some terms, Equation (1.3) becomes

$$f_Z(z;\gamma) = \frac{\gamma^{\gamma-1/2}\exp(-\gamma)}{\Gamma(\gamma)} \exp\Big(-\frac{z^2}{2} - \frac{z^3}{6\gamma^{1/2}} - \frac{z^4}{24\gamma} - ...\Big).$$

Using Stirling's approximation for the gamma function, which states

$$\Gamma(\gamma) = (2\pi)^{1/2}\gamma^{\gamma-\frac{1}{2}}\exp(-\gamma)\left(1 + O\left(\frac{1}{\gamma}\right)\right),$$

where the error term approaches 0 as $\gamma \to \infty$, and (1.3) becomes

$$f_Z(z;\gamma) \to \frac{\exp(-z^2/2)}{(2\pi)^{1/2}},$$

which is the desired limiting standard normal probability density function.

To relate Z back to $\log T$, we write

$$\begin{aligned} Z &= (W - \log\gamma)\sqrt{\gamma} \\ &= (\alpha\log(\lambda T) - \log\gamma)\sqrt{\gamma} \\ &= \frac{\log T - \left((\log\gamma)/\alpha - \log\lambda\right)}{1/(\alpha\sqrt{\gamma})}, \end{aligned} \tag{1.4}$$

and let $\mu_e = (\log\gamma)/\alpha - \log\lambda$, $\sigma_e = (\alpha\sqrt{\gamma})^{-1}$. These location and scale parameters μ_e and σ_e may differ from the μ and σ in a generalised F distribution; we will discuss this further in Chapter 4. With (1.4), we write

$$\log(T^*) = \mu_e + \sigma_e Z.$$

This shows that the log-normal distribution is a special case of the generalised gamma distribution when $\gamma \to \infty$.

It is useful to consider a further reparameterisation suggested by Prentice (1974) and Lawless (1982), which replaces γ with $Q = \gamma^{-1/2}$. This allows the case $Q < 0$ to be included in an extended family of log-gamma distributions. For these $Z = \frac{\log T^* - \mu_e}{\sigma_e}$, and the probability density function is

$$\frac{|Q|}{\Gamma(Q^{-2})}(Q^{-2})^{Q^{-2}} \exp\left[Q^{-2}(Qz - e^{Qz})\right].$$

Equivalently, it is easy to see that the density function for T^* is

$$f_{T^*}(t) = \frac{|Q|}{t\sigma_e\Gamma(Q^{-2})}(Q^{-2})^{Q^{-2}} \exp\left[Q^{-2}\left(Q\frac{\log t - \mu_e}{\sigma_e} - \exp\left(Q\frac{\log t - \mu_e}{\sigma_e}\right)\right)\right], \tag{1.5}$$

where $Q < 0$ indicates that $-Z$ has the probability density function $f_Z(z;\gamma)$, $Q = 0$ indicates that Z has a standard normal distribution, $Q > 0$ indicates that Z has the probability density function $f_Z(z;\gamma)$.

1.4.3 Proportional hazards model

In the context of survival analysis, it is also natural to consider the hazard rate, which is defined as

$$h(t) = \lim_{\mathrm{d}t\to 0} \frac{P(t \le T^* \le t + \mathrm{d}t \mid T^* \ge t)}{\mathrm{d}t} = \frac{F'(t)}{S(t)},$$

where F' is the derivative of the cumulative distribution for the survival time, assumed to exist, and S is the corresponding survival function. The hazard can be interpreted as the instantaneous probability of death for an individual at a particular time. In other words, if we define the cumulative hazard function as $H(t)$, then

$$h(t) = \mathrm{d}H(t) = H[t, t+\mathrm{d}t) = P\left(T^* \in [t, t+\mathrm{d}t) | T^* \ge t\right) = \frac{F'(t)}{S(t)} = \frac{\mathrm{d}}{\mathrm{d}t}\log\left(1 - F(t)\right).$$

Hence, the cumulative hazard function and the overall survival function can be expressed as

$$H(t) = \int_0^t h(u)\mathrm{d}u, \tag{1.6}$$

$$S(t) = 1 - F(t) = \exp\Big(-\int_0^t h(u)\mathrm{d}u\Big). \tag{1.7}$$

To determine the effect of covariates on the survival function and hazard rates, Cox (1972) proposed the famous proportional hazards assumption, such that for the ith individual with covariates $\mathbf{X}_i$, we write

$$h_i(t) = h_b(t)\exp(\boldsymbol{\beta}'\mathbf{X}_i),$$

where $h_b(t)$ is an arbitrary baseline hazard function, and $\boldsymbol{\beta}$ is the vector of coefficients for $\mathbf{X}_i$. Therefore, it is easy to see that the survival function can be written in the form

$$S_i(t) = S_b(t)^{\exp(\boldsymbol{\beta}'\mathbf{x}_i)},$$

where $S_b(t) = \exp\Big(-\int_0^t h_b(u)du\Big)$ is the baseline survival function.

Any parametric survival model can be easily transformed into a proportional hazards model by setting the baseline hazard function as the hazard function that corresponds to the model assumed. For example, a Weibull proportional hazards model can be implemented by setting the baseline hazard as

$$h_b(t) = \frac{\alpha\lambda(\lambda t)^{\alpha-1}\exp\{-(\lambda t)^\alpha\}}{\exp\{-(\lambda t)^\alpha\}} = \lambda^\alpha\alpha t^{\alpha-1}.$$

On the other hand, assuming a parametric lifetime distribution will restrict inferences to a family of models that can be indexed with a finite-dimensional set of parameters, and this process may sometimes over-simply the problem and cause misleading inference and decisions, as noted by Müller and Mitra (2013). To solve this problem, Cox suggested using a nonparametric lifetime distribution, namely the Kaplan-Meier estimator, which was defined in Theorem 1.1.

In this case, the baseline hazard is considered as nonparametric, while the hazard ratio $\exp(\boldsymbol{\beta}'\mathbf{X}_i)$ is modelled parametrically. This combination results in a semiparametric model, which is known as a Cox proportional hazards model. It is widely used in practice not only because it does not need a specific probabilistic distribution for the survival times, but also because the covariate effects on the hazard functions, $\boldsymbol{\beta}$, can be estimated separately using a

partial likelihood.

The parameters $\boldsymbol{\beta}$ can be estimated using maximum likelihood estimation, where the contribution of an observed failure to the overall likelihood, assuming there are no tied death times, is

$$\begin{aligned} L_j(\boldsymbol{\beta}) &= P\big(\textit{Case j fails}|\textit{There is 1 Failure From } R(T_j)\big) \\ &= \frac{P\big(\textit{Case j Fails} \mid \textit{Case j } \in R(T_j)\big)}{\sum_{l\in R(T_j)} P\big(\textit{Case l Fails} \mid \textit{Case l } \in R(T_j)\big)} \\ &= \frac{h_j(T_j)}{\sum_{l\in R(T_j)} h_l(T_l)} \\ &= \frac{\exp(\boldsymbol{\beta}'\mathbf{X}_j)}{\sum_{l\in R(T_j)} \exp(\boldsymbol{\beta}'\mathbf{X}_l)}, \end{aligned}$$

where $R(T_j)$ is the set of all cases at risk at time T_j.

Letting k be the total number of observed failures, we can obtain the estimators using the partial likelihood

$$L_p(\beta) = \prod_{j=1}^{k} \frac{\exp(\boldsymbol{\beta}'\mathbf{X}_j)}{\sum_{l\in R(T_j)} \exp(\boldsymbol{\beta}'\mathbf{X}_l)}. \tag{1.8}$$

Alternatively, we can first derive the complete likelihood given a set of survival data of size n, with notations as described in Section, as 1.3

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \Big[h_i(T_i)\exp\big(-H_i(T_i)\big)\Big]^{C_i}\Big[\exp\big(-H_i(T_i)\big)\Big]^{1-C_i} \\ &= \prod_{i=1}^{n} h_i(T_i)^{C_i}\exp\big(-H_i(T_i)\big) \\ &= \prod_{i=1}^{n} \Big[\frac{h_i(T_i)}{\sum_{l\in R(T_i)} h_l(T_l)}\Big]^{C_i}\Big[\sum_{l\in R(T_i)} h_l(T_l)\Big]^{C_i}\exp\big(-H_i(T_i)\big). \end{aligned}$$

Cox (1975) noted that the factor

$$\prod_{i=1}^{n} \Big[\sum_{l\in R(T_i)} h_l(T_l)\Big]^{C_i}\exp\big(-H_i(T_i)\big)$$

contains little information about $\boldsymbol{\beta}$, though it contains most of the information about the hazard function. Therefore, if the interest of the study is purely on the effect of covariates, we are able

to simplify the computation by only considering

$$\prod_{i=1}^{n} \Big[\frac{h_i(T_i)}{\sum_{l \in R(T_i)} h_l(T_l)} \Big]^{C_i}$$

in the complete likelihood, which is the partial likelihood as defined by Equation (1.8).

For the proportional hazards model, the covariate effects are obtained on the hazard functions, and the proportional hazards assumption is crucial. When this assumption is invalid, accelerated failure time models, whose covariate effects are obtained on (transformations of) the survival times, can be offered as alternatives for the proportional hazards models.

1.4.4 Accelerated failure time model

Similar to the proportional hazards models, different assumptions on the underlying survival function will result in either parametric or semi-parametric accelerated failure time models. In this section, we will briefly introduce the structure and properties of a general accelerated failure time model.

Recall from Section 1.3 that T_i^* denotes the true survival time, with distribution function F, and for the ith individual, $\mathbf{X}_i$ denotes its vector of covariate. The accelerated failure time model links the individual's log survival times to the covariates through a linear regression

$$\log T_i^* = \boldsymbol{\beta}' \mathbf{X}_i + \epsilon_i,$$

where the ϵ_i are independent and identically distributed random variables with $\exp(\epsilon_i)$ having some unspecified common distribution function F_ϵ. We will assume a log transformation on the true survival times throughout, although it can be replaced by any other strictly increasing function. We note that the intercept term β_0 usually included in the regression coefficient vector $\boldsymbol{\beta}$ is fixed to be 0 in an accelerated failure time model. Consequently, the mean of the error terms may not be zero. This is because, as noted in Wei (1992), in the presence of censoring, usually the intercept term will not be well estimated even if we include it in the regression.

We can express the accelerated failure time model in terms of the hazard function, where $h(t) = f(t)/S(t)$. Firstly, a simple transformation can be easily computed, where

$$\begin{aligned} F(t) &= P(T^* < t) \\ &= P(\boldsymbol{\beta}'\mathbf{X} + \epsilon < \log t) \\ &= P(\epsilon < \log t - \boldsymbol{\beta}'\mathbf{X}) \\ &= P\left(\exp(\epsilon) < t\exp(-\boldsymbol{\beta}'\mathbf{X})\right) \\ &= F_\epsilon\left(t\exp(-\boldsymbol{\beta}'\mathbf{X})\right). \end{aligned}$$

With the hazard function for $\exp(\epsilon)$ denoted as $h_\epsilon(t) = f_\epsilon(t)/S_\epsilon(t)$, where $f_\epsilon(t) = \frac{\mathrm{d}F_\epsilon(t)}{\mathrm{d}t}$ and $S_\epsilon(t) = 1 - F_\epsilon(t)$, we are able to express the hazard function for T^* as

$$\begin{aligned} h(t) &= \frac{\mathrm{d}F(t)}{\mathrm{d}t}/(1 - F(t)) \\ &= \frac{\frac{\mathrm{d}}{\mathrm{d}t}F_\epsilon\left(t\exp(-\boldsymbol{\beta}'\mathbf{X})\right)}{1 - F_\epsilon\left(t\exp(-\boldsymbol{\beta}'\mathbf{X})\right)} \\ &= \frac{f_\epsilon\left(t\exp(-\boldsymbol{\beta}'\mathbf{X})\right)\exp(-\boldsymbol{\beta}'\mathbf{X})}{1 - F_\epsilon\left(t\exp(-\boldsymbol{\beta}'\mathbf{X})\right)} \\ &= h_\epsilon\left(t\exp(-\boldsymbol{\beta}'\mathbf{X})\right)\exp(-\boldsymbol{\beta}'\mathbf{X}). \end{aligned}$$

From the above result, it is clear that the hazards are not proportional in the accelerated failure time framework. The accelerated failure time model enables us to interpret the baseline covariates as either decelerating or accelerating the event of interest. Furthermore, while the proportional hazards structure does not provide a helpful parametric distinction between the effects of the covariates on the timing of the event, the accelerated failure time structure allowed this distinction, which is desired in some practical situations.

The effect of the covariate on the "baseline survival function" $S_\epsilon(t)$, which is the survival function of $\exp(\epsilon)$, can be represented as

$$\begin{aligned} S(t) &= \exp\left(-\int_0^t h(u)du\right) \\ &= \exp\left(-\int_0^t h_\epsilon\left(u\exp(-\boldsymbol{\beta}'\mathbf{X})\right)\exp(-\boldsymbol{\beta}'\mathbf{X})du\right) \\ &= \exp\left(-H_\epsilon\left(t\exp(-\boldsymbol{\beta}'\mathbf{X})\right)\right) \\ &= S_\epsilon\left(t\exp(-\boldsymbol{\beta}'\mathbf{X})\right), \end{aligned}$$

where H_ϵ denotes the cumulative hazard function for $\exp(\epsilon)$.

1.5 Thesis structure and contribution

The structure of the rest of this thesis is as follows.

Chapter 2 starts with an in-depth review of the construction, development and application of some of the most commonly used cure models in practice. Though in this thesis we focus on mixture cure models, which will be introduced in Section 2.2, we introduce and discuss some of the non-mixture cure models in Section 2.3.

After that, we discuss the issue of insufficient follow-up in Chapter 3. We start by describing the "near non-identifiability problem" faced by parametric mixture cure models when the follow-up time is insufficient in Section 3.2. Then we review some existing methods and hypothesis tests to identify insufficient follow-up in Section 3.3.

Among all models introduced, the researchers will need to select an appropriate model to use in practice. Using the parametric mixture cure models as examples, a starting point for model selection would be to find a single parametric model, which embeds the most suitable parametric lifetime distributions. We therefore introduce the generalised F distribution in Section 4.2, which includes the generalised gamma distribution and the generalised log-logistic distributions and many others as its sub-models. However, to the best of our knowledge, the connections between the generalised F distribution and its sub-models have never been clearly delineated in the literature. This motivates us to construct such relationships in detail in Section 4.3. With these results, we can then perform parametric mixture cure model selection using the generalised F distribution. We explain and demonstrate model selection with examples in Chapter 5. In Section 5.5, we apply the model selection process to real cancer datasets and conduct statistical inferences using parametric mixture cure models. After conducting tests for sufficient follow-up in Section 5.5.1, we demonstrate the computations involved in the model selection process in Section 5.5.2. Lastly, we demonstrate the methodologies for covariate analysis in Section 5.5.3.

Chapter 6 covers topics in cure proportion estimation under both sufficient and insufficient follow-up. We review both parametric and nonparametric cure proportion estimators under sufficient follow-up in Section 6.2. However, we note that they will not perform well under

insufficient follow-up; for example, the nonparametric cure proportion estimator always overestimates the cure proportion under insufficient follow-up. We then review a nonparametric cure proportion estimator proposed by Escobar-Bach and Van Keilegom (2019) in Section 6.3, which uses extrapolation techniques from extreme value theory to offset bias under insufficient follow-up. Their estimator applies when the survival distribution of those susceptible to the event belongs to the Fréchet maximum domain of attraction. Besides the Fréchet, an important class of limiting distributions for maxima is the Gumbel class. We show in Section 4.4 that a vast class of commonly used survival distributions, the generalised Gamma distributions, are in the Gumbel domain of attraction. In Sections 6.4 and 6.5 we use extreme value theory to derive, for distributions in this class, a nonparametric estimator of the cure proportion that is consistent and asymptotically normally distributed under reasonable assumptions, and performs well in simulation studies with data where follow-up is insufficient. We illustrate its use with an application to survival data where patients with differing stages of breast cancer have varying degrees of follow-up in Section 6.6.

Chapter 2

Common Cure Models Applied in Practice

2.1 Introduction

Section 1.4 reviews only the traditional survival models, and apart from the nonparametric Kaplan-Meier estimator, the other models cannot be directly applied to data with long term survivors. To analyse such data, Boag introduced mixture cure models into the field of survival analysis in 1949. These models have attracted attention over the past few decades, and an extensive literature has developed. We refer to Maller and Zhou (1996), who treat cure models in a systematic way with a primary focus on mixture cure models.

Othus, Barlogie, Leblanc and Crowley (2012) reviewed the structural differences between standard survival models and various cure models, with applications to a multiple myeloma dataset. From their work, it was clear that the numerical results significantly differ, implying that applying standard survival models may lead to incorrect inferences when the data contains immunes. Therefore, they "hope that some researchers will find the (cure) models a useful alternative to standard survival models when analysing some types of cancer survival data".

While the targeted readers for the review paper by Othus et al. (2012) are medical researchers without solid statistical backgrounds, Taweab and Ibrahim's (2014) review paper is more formal mathematically. It summarised the development of cure models until 2014, with a particular focus on the mixture and non-mixture cure models. Variations of the mixture and non-mixture cure models are briefly mentioned in their paper but not described in detail. In addition, they included a brief discussion on the "change-point" problem, where the covariates may have differing effects on the hazard rate after a threshold time.

Peng and Taylor (2014) wrote a chapter in *Handbook of Survival Analysis* (Klein, John P., Hans C. Van Houwelingen, Joseph G. Ibrahim, and Thomas H. Scheike, 2014), reviewing the concept of cure models in survival analysis. It explained the model formulation, estimation methods and applications for various mixture cure models, the non-mixture cure models and some other special cure models. The authors also discussed the identifiability issue for some of the cure models reviewed and its effects on inference, especially for cure proportion estimation. Amico and Van Keilegom (2018) produced an updated review of cure models.

In this chapter, we will first review some of the commonly used mixture and non-mixture cure models by introducing their formulation, estimation methods, and strengths and weaknesses in Sections 2.2 and 2.3.

2.2 Mixture Cure Models

The original idea of a mixture cure model is due to Boag (1949). He proposed a log-normal distribution to model the failure time of the susceptibles and assumed the cure probability to be constant. The cured proportion was then estimated using maximum likelihood. The model was further developed by Berkson and Gage (1952) using a life-table approach. In the binary case where the only covariate X is a Bernoulli variable with coefficients β, and l_t is the proportion of the total population surviving to time t, they write

$$l_t = (1-p)l_0 + pl_0 \exp(-\beta t),$$

where l_0 is the net survivorship corresponding to "normal" deaths, obtained from standard life tables. The Berkson and Gage (1952) model is structurally similar to the mixture cure models introduced by Boag (1949).

The structure for a mixture cure model that will be used throughout this thesis is defined in Definition 2.1.

Definition 2.1 (Mixture cure model)**.** Using the notation defined in Section 1.3, we note the distribution for T is in general improper, but the distribution of the susceptibles' survival time, T^*, is proper (the distribution function $F_0(t) \to 1$ as $t \to \infty$). With the introduction of independent latent Bernoulli random variables B_i, where $P(B_i = 1) = p$ with $p \in [0,1]$, we are able to relate

F to F_0 as follows

$$\begin{aligned} F(t) &= P(T_i^* \leq t) \\ &= P(T_i^* \leq t | B_i = 1) \times P(B_i = 1) + P(T_i^* \leq t | B_i = 0) \times P(B_i = 0) \\ &= pF_0(t),\ t > 0. \end{aligned} \tag{2.1}$$

Equivalently, the survival function of a mixture cure model for $t \leq T^+$, where T^+ is an arbitrary large time beyond which we have no interest, can be expressed as

$$S(t) = 1 - p + pS_0(t).$$

Most of the relevant literature favours the use of the mixture cure models because they are conceptually and computationally simple; they are easy to interpret; and they can be generalised to more complex situations, as outlined in Peng and Taylor (2014). At the same time, they are flexible when modelling the covariate effects on the event time distribution. Othus et al. (2012) have also noted that the mixture cure models allow the covariates to have a different effect on immune individuals and susceptible individuals, which can be regarded as another advantage.

As suggested by many authors including Farewell (1982), we are able to fit a model $p(\mathbf{X})$ for the susceptible proportion depending on the covariates $\mathbf{X}$, and another separate model with distribution function $F_0(t; \mathbf{X})$ for the susceptible survival time. The resulting model for $p(\mathbf{X})$ is referred to as "the incidence model", whose vector of regression coefficients is denoted as $\mathbf{b}$. The model for $F_0(t; \mathbf{X})$ is referred to as "the latency model", whose vector of regression parameters is denoted as $\boldsymbol{\beta}$. Mixture cure models differ by making different assumptions for the incident model and the latency model. For example, if we assume a parametric distribution for the latency model, the overall cure model will become a parametric mixture cure model. Common choices for the parametric latency model include an exponential distribution, Weibull distribution, log-normal distribution or generalised gamma distribution. Alternatively, if a semi-parametric distribution is assumed for the latency model, the overall cure model becomes a semi-parametric mixture cure model. For the incident model, most mixture cure models use logistic regression because it is simple to compute and interpret.

2.2.1 Identifiability of Mixture Models

When analysing survival data with censored observations, one common assumption that is often made implicitly is that survival and censoring times are independent. If they are related, the censoring is informative. The results in Cox (1959), Tsiatis (1975) and Williams and Lagakos (1977) showed that from the observed data, it is impossible to distinguish a model whose censoring is informative from one whose censoring is non-informative. Williams and Lagakos (1977) and Lagakos (1979) claimed the only means of overcoming this identifiability problem is to make specifications and restrictions on the family of survival model. We refer to Elandt-Johnson and Johnson (1980) and Contant Jr (1988) for comprehensive descriptions of this identifiability problem.

Lagakos (1979) listed three examples of situations where the censoring is possibly informative, with the first being "a clinical trial in which some patients remove themselves from study for reasons possibly related to therapy and thereby censor their survival time under test conditions". Therefore, when patients are removed from study because they are cured by the therapy (i.e. there are immunes in the population), the censoring is informative. In order to overcome the identifiability problem, we need to consider making specifications and restrictions on the cure models.

The identifiability of mixture cure models is studied by Li, Taylor and Sy (2001) and Hanin and Huang (2014). Let $\mathcal{F}_0 = \{F_0(t;\boldsymbol{\theta}|X) : \boldsymbol{\theta} \in \Theta\}$ be the class of proper lifetime distributions for the susceptible, and $\mathcal{X}$ be the design space. Let $\mathcal{P} = \{p(X) : p(X) \neq 0 \text{ or } 1, \text{for all } X \in \mathcal{X}\}$ denote the space of incident models, where the authors excluded functions $p(X)$ that define degenerate mixtures. With the inclusion of covariates, without loss of generosity, the authors assumed that X is one-dimensional, hence X is the closed interval $[a_0, a_1]$. Therefore, the class of mixture cure models can be defined by:

$$\mathcal{F} = \{F(t;\boldsymbol{\theta},X) : F(t;\boldsymbol{\theta},X) = p(X)F_0(t;\boldsymbol{\theta}|X), X \in \mathcal{X}, p(\cdot) \in \mathcal{P}, F_0(\cdot;\boldsymbol{\theta}|X) \in \mathcal{F}_0\}.$$

Following the approach of Titterington et al. (1985), Li, Taylor and Sy (2001) defined the identifiability of the cure models (both mixture and non-mixture) as follows.

Definition 2.2 (Definition 1 in Li, Taylor and Sy (2001))**.** The class of mixture cure models $\mathcal{F}$ is identifiable if for any two members of $\mathcal{F}$ given by $F(t;\boldsymbol{\theta},X) = p(X)F_0(t;\boldsymbol{\theta}|X)$ and $F_{(1)}(t,\boldsymbol{\theta}_{(1)},X) = p_{(1)}(X)F_{0,(1)}(t;\boldsymbol{\theta}_{(1)}|X)$, then the function $F(t;\boldsymbol{\theta},X)$ is identical to $F_{(1)}(t;\boldsymbol{\theta}_{(1)},X)$ if and only if

$p(X)$ is identical to $p_{(1)}(X)$ for $X \in \mathcal{X}$ and $F_0(t;\boldsymbol{\theta})|X)$ is identical to $F_{0,(1)}(t;\boldsymbol{\theta}_{(1)}|X)$ for almost all $t > 0$.

We now discuss the identifiability of mixture cure models.

Theorem 2.1 (Theorem 3 in Li, Taylor and Sy (2001))**.** *Under regularity conditions, for a mixture cure model as defined by Definition 2.1, when the parameters for F_0 have a finite dimension (i.e. F_0 is assumed to be parametric), and X is a continuous one-dimensional covariate in the design space $\mathcal{X} = [a_0, a_1]$, where $-\infty < a_0 < a_1 < \infty$ (i.e. covariate is continuous and bounded), the model is identifiable regardless of whether $p(X)$ is parametric or not.*

Apart from the parametric assumptions for the latency model, some users prefer a nonparametric or semiparametric assumption to let the data speak instead of a subjective model. On the other hand, if this is assumed in practice, the incident model must have a parametric structure; otherwise, the overall mixture cure model may be unidentifiable without including covariates by Theorem 2.1. If a logistic regression is assumed for the incident model, the resulting mixture cure model consists of a parametric incident model and a nonparametric or semiparametric latency model. This combination results in a semiparametric mixture cure model. For example, if we assume a Cox proportional hazards model, which was introduced in Section 1.4.3, for the latency model and logistic regression for the incident model, the overall model would be a semiparametric mixture cure model, which is identifiable under some conditions that are outlined in Theorem 2.2.

Theorem 2.2 (Theorem 2 in Li, Taylor and Sy (2001).)**.** *Under regularity conditions, for a mixture cure model as defined by Definition 2.1, when F_0 is assumed to have a proportional hazards structure, and X is a continuous one-dimensional covariate in the design space $\mathcal{X} = [a_0, a_1]$, where $-\infty < a_0 < a_1 < \infty$ (i.e. covariate is continuous and bounded), the model is identifiable if the relative hazard function $r_h(X)$ (usually specified as $\exp(\boldsymbol{\beta}'X)$) is defined on $\mathcal{R} = \{r_h(X) : r_h(X) > 0$, $r_h(X)$ is a non-constant function, $r_h(0) = 0\}$.*

However, according to Theorem 2.3, without considering covariates, the overall mixture cure model suffers from identifiability issues if both the incident and latency models are unspecified or assumed nonparametric. Therefore, we do not review nonparametric mixture cure models in this section; however, we refer to López-Cheda et al. (2017).

Theorem 2.3 (Theorem 1 in Li, Taylor and Sy (2001).)**.** *Under regularity conditions, for a mixture cure model as defined by Definition 2.1, when the parameters for F_0 do not have a finite dimension (i.e. F_0 is assumed to be nonparametric), and X is a continuous one-dimensional covariate in the design space $\mathcal{X} = [a_0, a_1]$, where $-\infty < a_0 < a_1 < \infty$ (i.e. covariate is continuous and bounded):*

1. *the mixture cure model is not identifiable if $p(X)$ is unspecified, unless $S_0(t^+) = 0$ and $\max_{X \in \mathcal{X}} p(X) = 1$.*
2. *the mixture cure model is not identifiable if $p(X)$ is a constant parameter p.*
3. *the mixture cure model is identifiable if $p(X)$ is specified as a logistic function (2.2).*

In the following subsections, we will list a few of the mixture models that have been widely used in practice.

2.2.2 Weilbull Mixture Cure Model

Farewell (1982) formally introduced mixture cure models to survival analysis and extended the model by adding in logistic regression for the dependence of the cured proportion on covariates $\mathbf{X}$, so that

$$p(\mathbf{X}) = \frac{\exp(\mathbf{b}'\mathbf{X})}{1 + \exp(\mathbf{b}'\mathbf{X})}, \tag{2.2}$$

where $\mathbf{b}$ is the vector of regression coefficients in the incidence model. For the latency model, Farewell (1982) assumed a Weilbull distribution, specifically,

$$\begin{aligned} S_0(t) &= \exp(-\lambda t)^{\alpha}, \\ f_0(t) &= \alpha\lambda(\lambda t)^{\alpha-1} \exp\{-(\lambda t)^{\alpha}\}, \end{aligned}$$

where λ and α are the scale and shape parameters for the Weibull distribution. With specifications for both the latency and the incident models, a parametric mixture cure model is defined by Definition 2.1. Hence, given a dataset $\mathcal{D} = \{T, C, \mathbf{X}\}$, we can compute the overall likelihood function as

$$L(\underset{\sim}{\theta}; \mathcal{D}) = \prod_{i=1}^{n} \left[p(\mathbf{X}_i; \mathbf{b}) f_0(T_i; \alpha, \lambda)\right]^{C_i} \left[1 - p(\mathbf{X}_i; \mathbf{b}) + p(\mathbf{X}_i; \mathbf{b}) S_0(T_i; \alpha, \lambda)\right]^{1-C_i},$$

where $\underset{\sim}{\theta} = \{\alpha, \lambda, \mathbf{b}\}$ is the set of parameters.

Farewell (1982) suggested using Newton-Raphson iteration to obtain the maximum likelihood estimators for the parameters $\underset{\sim}{\theta}$. Nowadays we can use programming languages such as R to perform the Newton-Raphson algorithm. For example, we will demonstrate using the "flexsurvcure" package in R for computation in Section 5.4.1. In addition, to account for the effect of covariates in the latency model, we can express the parameters of the Weibull distribution in terms of $\exp(\boldsymbol{\beta}'\mathbf{X})$ and again use maximum likelihood to estimate the coefficients $\boldsymbol{\beta}$. For example, the scale parameter can be expressed in the form

$$\log \lambda = \boldsymbol{\beta}_\lambda' \mathbf{X}.$$

As mentioned in Section 1.4.2, apart from the Weibull distribution, there are numerous other parametric lifetime distributions that can be utilised in survival analysis. Ideally, they can all be applied as parametric assumptions for the latency model in the cure model context, with very similar estimation methods. Examples include the exponential mixture cure model and the log-normal mixture cure model. We note that according to Theorem 2.1, such parametric mixture cure models are identifiable.

2.2.3 Generalised Gamma Mixture Accelerated Failure Time Cure Model

Yamaguchi (1992) adapted the structure of mixture cure models to the accelerated failure time model framework. Specifically, the survival times for the susceptibles are modelled by

$$\log T^*|_{B=1} = \boldsymbol{\beta}'\mathbf{X} + \sigma Z = \mu + \sigma Z,$$

where σ is a scale parameter and Z has a standard log-gamma distribution with shape parameter γ, as shown in Equation (1.3), such that

$$f_Z(z;\gamma) = \begin{cases} \frac{\gamma^{\gamma-1/2}}{\Gamma(\gamma)} \exp\left(\sqrt{\gamma}z - \gamma e^{z/\sqrt{\gamma}}\right), & \text{if } 0 \leq \gamma < \infty, \\ 1, & \text{if } \gamma = \infty. \end{cases}$$

The log-Gamma model and extended generalised gamma distribution were discussed in Section 1.4.2.

For the parametric accelerated failure time cure model, the dependence of the susceptible proportion p on the covariates is again modelled by a logistic regression in the same form as

shown by Equation (2.2).

Given the observed data $\mathcal{D} = \{T, C, \mathbf{X}\}$, let

$$Z = \frac{\log T^* - \boldsymbol{\beta}'\mathbf{X}}{\sigma},$$

so that the covariate effects are modelled through $\mu = \boldsymbol{\beta}'\mathbf{X}$. If the i-th observation is uncensored, so $C_i = 1$, its contribution to the overall likelihood function is

$$p(\mathbf{X}_i) f_0(T_i) = p(\mathbf{X}_i) f_Z(z_i; \gamma, \boldsymbol{\beta}, \sigma) \left| \frac{\mathrm{d}Z}{\mathrm{d}\log T^*} \right| = \frac{p(\mathbf{X}_i) f_z(z_i; \gamma, \boldsymbol{\beta}, \sigma)}{\sigma T_i}. \tag{2.3}$$

On the other hand, if the observation is censored, so $C_i = 0$, its contribution will be

$$1 - p(\mathbf{X}_i) + p(\mathbf{X}_i) S_0(\log T_i^*).$$

To evaluate the survival function for the susceptibles, we again apply the simple transformation using the relationship (2.2.3) and the known probability density function for Z (1.3), where

$$\begin{aligned} S_0(\log T_i^*) &= 1 - F_0(\log T_i^*) \\ &= 1 - F_Z(z_i) \\ &= \int_{z_i}^{\infty} \frac{\gamma^{\gamma - 1/2}}{\Gamma(\gamma)} \exp\left(\sqrt{\gamma} u - \gamma \exp(u/\sqrt{\gamma})\right) \mathrm{d}u. \end{aligned}$$

Consider a simple change of variable

$$m = \gamma \exp(u/\sqrt{\gamma}),$$

so that

$$\begin{aligned} u &= \sqrt{\gamma} \log(m/\gamma), \\ \mathrm{d}m &= \sqrt{\gamma} \exp(u/\sqrt{\gamma}) \mathrm{d}u. \end{aligned}$$

Then we write

$$\begin{aligned} S_0(\log T_i^*) &= \int_m^\infty \frac{\gamma^\gamma}{\sqrt{\gamma}\Gamma(\gamma)} \exp\left(\sqrt{\gamma}u\right) \exp(-m) \frac{\mathrm{d}m}{\sqrt{\gamma}\exp(u/\sqrt{\gamma})} \\ &= \int_m^\infty \frac{\gamma^\gamma}{\sqrt{\gamma}\Gamma(\gamma)} \exp(\gamma \log(m/\gamma)) \exp(-m) \frac{\mathrm{d}m}{m/\sqrt{\gamma}} \\ &= \int_m^\infty \frac{\gamma^\gamma}{\sqrt{\gamma}\Gamma(\gamma)} \frac{m^\gamma}{\gamma^\gamma} \exp(-m) \frac{\gamma}{\sqrt{\gamma}} \frac{\mathrm{d}m}{m} \\ &= \int_m^\infty \frac{m^{\gamma-1}}{\Gamma(\gamma)} \exp(-m)\mathrm{d}m \\ &= 1 - I\left(\gamma, \gamma\exp(z_i/\sqrt{\gamma})\right), \end{aligned} \tag{2.4}$$

where $I(\gamma, a) = 1 - \int_a^\infty \frac{m^{\gamma-1}}{\Gamma(\gamma)} \exp(-x)\mathrm{d}x$ is the lower incomplete gamma integral as defined in (1.4.2). Therefore, combining the results from (2.3) and (2.4), the complete likelihood given data $\mathcal{D}$ is

$$L(\mathbf{b}, \boldsymbol{\beta}, \gamma, \sigma) = \prod_{i=1}^n \left(\frac{p_i f_Z(z_i; \gamma, \beta, \sigma)}{\sigma T_i} \right)^{C_i} \left(1 - p_i + p_i \left(1 - I\left(\gamma, \gamma\exp(z_i/\sqrt{\gamma})\right)\right)\right)^{1-C_i},$$

for $0 < \gamma < \infty$.

Yamaguchi (1992) further extended and generalised the complete likelihood by introducing a new parameter $\lambda_\gamma = \gamma^{-1/2}$. This idea was adopted from Prentice (1974), with details given in Section 1.4.2. The practical method of this extension can be categorized into three situations. Firstly, when $\gamma = \infty$, Yamaguchi (1992) suggested replacing $1 - I\left(\gamma, \gamma\exp(z_i/\sqrt{\gamma})\right)$ by the normal integral from z_i to infinity. Secondly, for positive λ_γ, simply replace γ by λ_γ^{-2}. Lastly, for negative λ_γ, replace $f_Z(z_i; \gamma, \beta, \sigma)$ by $f_Z(z_i; \lambda_\gamma^{-2}, \beta, \sigma)$ and $1 - I\left(\gamma, \gamma\exp(z_i/\sqrt{\gamma})\right)$ by $I\left(\lambda_\gamma^{-2}, \lambda_\gamma^{-2}\exp(-z_i|\lambda_\gamma|)\right)$.

We note that according to Theorem 2.1, the mixture accelerated failure time models introduced by Yamaguchi are all identifiable. Though Yamaguchi chose to assume the generalised and the extended log-gamma model for the random error Z, they are not the only possible parametric assumptions. In fact, different assumptions will lead to different mixture accelerated failure time cure models. They are all applicable in practice if the researchers believe the covariates are either decelerating or accelerating the event of interest for their susceptible population. For example, Peng, Dear and Denham (1998) assumed a generalised F distribution for Z, which is another viable option.

2.2.4 Semiparametric Cox Proportional Hazards Mixture Cure Model

As previously mentioned, mixture cure models differ from each other in the different assumptions for the latency model. It is natural to assume a Cox proportional hazards model for the susceptibles' survival times in a mixture cure model framework since it is arguably the most widely used survival model. Kuk and Chen (1992) first implemented this idea and utilised the marginal likelihood to estimate the parameters of interest. Sy and Taylor (2000) adopted similar ideas but simplified the computations by introducing the Expectation-Maximisation (EM) method to estimate the parameters. Peng and Dear's (2000) paper presented similar content to Sy and Taylor (2000).

In Section 1.4.3, we introduced and discussed the proportional hazards model in detail. In this section, we define $h_b(t|B=1)$ as the baseline hazard function for the susceptible and $S_b(t|B=1)$ as the baseline survival function; then given the covariates vector $\mathbf{X}$ and the coefficient vector $\boldsymbol{\beta}$, for each individual i,

$$h_0(T_i) = h_b(T_i|B_i=1)\exp(\boldsymbol{\beta}'\mathbf{X}_i),$$
$$S_0(T_i) = S_b(T_i|B_i=1)^{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}.$$

When Kuk and Chen (1992) first introduced the Cox proportional hazards mixture cure model, they used marginal likelihood for parameter estimation. Given a set of survival data $\mathcal{D} = \{T, C, \mathbf{X}\}$, let $\tau_{(1)} < \tau_{(2)} < \cdots < \tau_{(k)}$ denote the ordered uncensored observations, with $\mathbf{X}_{(1)}, ..., \mathbf{X}_{(k)}$ as the corresponding covariates, and further define $\tau_{(0)} = 0$ and $\tau_{(k+1)} = \infty$. Let $D := \{i; C = 1\}$ be the set of labels corresponding to the uncensored observations, and $C(j)$ indicate the set of labels corresponding to those individuals censored between the interval $\left[\tau_{(j)}, \tau_{(j+1)}\right)$, i.e. $l \in C(i)$ are the labels assigned for the censored observations with censoring times in $\left[\tau_{(i)}, \tau_{(i+1)}\right)$. The probability function is given by

$$\int_{\tau_{(1)}<\cdots<\tau_{(k)}} \cdots \int \prod_{i=1}^{k} \left[p(\mathbf{X}_{(i)}) f_0(\tau_{(i)}|\mathbf{X}_{(i)}) \prod_{l\in C(i)} \{1 - p(\mathbf{X}_l) + p(\mathbf{X}_l) S_0(\tau_{(i)}|\mathbf{X}_l)\} \right] \mathrm{d}\tau_{(k)}, \cdots, \mathrm{d}\tau_{(1)}, \quad (2.5)$$

assuming censoring can only occur immediately after failures and there are no ties. Kuk and Chen (1992) further rewrote the marginal likelihood (2.5) as

$$L(\mathbf{b}, \boldsymbol{\beta}) = \sum_{B_c \in \Omega} L(\mathbf{b}, \boldsymbol{\beta}; B_c), \tag{2.6}$$

where Ω is the collection of all n_c-tuples of 0's and 1's, and n_c is the number of censored observations. The set $B_c \in \Omega$ is a realization of the set of B_i with its labels $i \in C_0$, where $C_0 := \{i : C_i = 0\}$ is the set of labels assigned to censored observations. The components of (2.6) are defined as

$$L(\mathbf{b}, \boldsymbol{\beta}; B_c) = \Big(\prod_{i \in D} p(\mathbf{X}_{(i)})\Big) \prod_{i \in C_0} \Big\{p(\mathbf{X}_{(i)})^{B_l}(1 - p(\mathbf{X}_{(i)}))^{1-B_i}\Big\} \prod_{i=1}^{k} \Big[\psi_{(i)} / \Big\{\sum_{i=1}^{k}(\psi_{(j)} + \sum_{l \in C_{(j)}} B_l \psi_l)\Big\}\Big],$$

where $\psi_{(i)} = \exp(\boldsymbol{\beta}' \mathbf{X}_{(i)})$ and $\psi_l = \exp(\boldsymbol{\beta}' \mathbf{X}_l)$. Estimation of the regression parameters is by maximizing the marginal likelihood $L(\mathbf{b}, \boldsymbol{\beta})$.

The above marginal likelihood approach is not widely applied in practice. As an alternative, the Expectation-Maximisation algorithm in cure models was proposed by Sy and Taylor (2000) to simplify the computations and it was better accepted by researchers. Furthermore, utilising the Expectation-Maximisation algorithm also enables us to maximise the likelihoods for the parameters of the incident and the latency model ($\mathbf{b}$ and $\boldsymbol{\beta}$) separately. In fact, the Expectation Maximisation algorithm arises naturally since we introduced the latent Bernoulli random variable B, which is not observed. The detailed steps involved in the algorithm will be outlined as follows.

Assuming B, the susceptible indicator, is part of the complete data but unobserved, with $p_i = p(\mathbf{X}_i)$ and recalling the relationship (1.4.3), we write the complete likelihood function as

$$\begin{aligned} L_C(\mathbf{b}, \beta, h_b(t|B=1); B, \mathbf{X}) &= \prod_{i=1}^{n} \Big[\big\{p_i h_0(T_i|\mathbf{X}_i) S_0(T_i|\mathbf{X}_i)\big\}^{B_i}\Big]^{C_i} \times \Big[(1-p_i)^{1-B_i}\big(p_i S_0(T_i|\mathbf{X}_i)\big)^{B_i}\Big]^{1-C_i} \\ &= \prod_{i=1}^{n} p_i^{B_i}(1-p_i)^{(1-B_i)(1-C_i)} \prod_{i=1}^{n} h_0(T_i|\mathbf{X}_i)^{C_i B_i} S_0(T_i|\mathbf{X}_i)^{B_i} \\ &:= L_1(\mathbf{b}; B, \mathbf{X}) \times L_2(\boldsymbol{\beta}; B, \mathbf{X}). \end{aligned}$$

We can simplify the likelihood function $L_1(\mathbf{b}; B, \mathbf{X})$ by first reordering the individuals in the dataset by their survival times and censoring indicators. Let $\{(1), (2), ..., (n)\}$ be the set of indices for the reordered dataset, so that $\tau_{(1)} < \tau_{(2)} < \cdots < \tau_{(k)}$ are the survival times for the

uncensored observations, and $\tau_{(k+1)} < \cdots < \tau_{(n)}$ are the censored times for the censored observations. Let the corresponding vector of covariates for $\tau_{(i)}$ be indexed as $\mathbf{X}_{(i)}$, the corresponding censoring indicator for $\tau_{(i)}$ be indexed as $c_{(i)}$, and the corresponding susceptible indicator for $\tau_{(i)}$ be indexed as $B_{(i)}$. Then, with $p_{(i)} = p(\mathbf{X}_{(i)})$,

$$\begin{aligned} L_1(\mathbf{b}; B, \mathbf{X}) &= \prod_{i=1}^{n} p_i^{B_i} (1-p_i)^{(1-B_i)(1-C_i)} \\ &= \prod_{i=1}^{k} p_{(i)}^{B_{(i)}} (1-p_{(i)})^{(1-B_{(i)})(1-c_{(i)})} \prod_{i=k+1}^{n} p_{(i)}^{B_{(i)}} (1-p_{(i)})^{(1-B_{(i)})(1-c_{(i)})} \\ &= \prod_{i=1}^{k} p_{(i)}^{B_{(i)}} (1-p_{(i)})^{0} \prod_{i=k+1}^{n} p_{(i)}^{B_{(i)}} (1-p_{(i)})^{(1-B_{(i)})} \\ &= \prod_{i=1}^{n} p_i^{B_i} (1-p_i)^{(1-B_i)}. \end{aligned} \tag{2.7}$$

Note that when $1 < i < k$, $c_{(i)} = 1$ almost surely, which implies $B_{(i)} = 1$. The log-likelihood functions can therefore be expressed as

$$\begin{aligned} l_1(\mathbf{b} \mid B, \mathbf{X}) &= \sum_{i=1}^{n} B_i \log p_i + (1-B_i) \log(1-p_i), \\ l_2(\boldsymbol{\beta} | B, \mathbf{X}) &= \sum_{i=1}^{n} B_i C_i \log \left[h_0(T_i | \mathbf{X}_i) \right] + B_i \log \left(S_0(T_i | \mathbf{X}_i) \right). \end{aligned}$$

The Expectation-Maximisation algorithm consists of two steps. The "Expectation step" computes the expectation of the complete log-likelihood $l_C(b, \beta, h_b(t|B=1); B, \mathbf{X})$ with respect to the unobserved B given the current parameter values at the $(m+1)$-th iteration, which are $\Theta^{(m)} := \{\mathbf{b}^{(m)}, \boldsymbol{\beta}^{(m)}, h_b^{(m)}(t|B=1)\}$, and the observed data, which is $O = \{\text{observed } B, T, C, \mathbf{X}\}$. For the first iteration, we use some arbitrary values for $\Theta^{(0)}$ as the starting values. For the "observed" values for B_i required in each iteration, we use the conditional expectation of B_i. Note that for uncensored observation i,

$$\mathbb{E}[B_i | \theta^{(m)}, O, C_i = 1] = 1. \tag{2.8}$$

For censored individuals, both $l_1(\mathbf{b}|B,\mathbf{X})$ and $l_2(\boldsymbol{\beta}|B,\mathbf{X})$ are linear functions of B_i,

$$\begin{aligned} w_{i,C_i=0}^{(m)} &= \mathbb{E}[B_i|\Theta^{(m)}, O, C_i = 0] \\ &= P(B_i = 1|\Theta^{(m)}, C_i = 1) \\ &= \frac{p_i S_0(T_i|\mathbf{X}_i)}{1 - p_i + p_i S_0(T_i|\mathbf{X}_i)}\Big|_{\Theta^{(m)}}. \end{aligned}$$

The above function is only for the censored observations, and it represents the fractional allocation of the overall conditional expectation of B_i to the susceptible group, or in other words, the conditional probability that the individual will experience the event when $C_i = 0$. To include the uncensored cases with (2.8), we write

$$w_i^{(m)} = \mathbb{E}[B_i|\Theta^{(m)}, O] = C_i + (1 - C_i)\frac{p_i S_0(T_i|\mathbf{X}_i)}{1 - p_i + p_i S_0(T_i|\mathbf{X}_i)}\Big|_{\Theta^{(m)}}. \tag{2.9}$$

After this, for the $(m+1)$-th iteration, we replace the B_i's with $w_i^{(m)}$ to compute the conditional expectations of $l_1(\mathbf{b}|B,\mathbf{X})$ and $l_2(\boldsymbol{\beta}|B,\mathbf{X})$, where

$$\mathbb{E}[l_1(\mathbf{b}|B,\mathbf{X})] = \sum_{i=1}^{n} w_i^{(m)} \log(p_i) + (1 - w_i^{(m)}) \log(1 - p_i) \tag{2.10}$$

$$\begin{aligned} \mathbb{E}[l_2(\boldsymbol{\beta}|B,\mathbf{X})] &= \sum_{i=1}^{n} w_i^{(m)} C_i \log\left[h_0(T_i|\mathbf{X}_i)\right] + w_i^{(m)} \log\left[S_0(T_i|\mathbf{X}_i)\right] \\ &= \sum_{i=1}^{n} 0 + w_i^{(m)} C_i \log\left[h_0(T_i|\mathbf{X}_i)\right] + w_i^{(m)} \log\left[S_0(T_i|\mathbf{X}_i)\right] \\ &= \sum_{i=1}^{n} C_i \log w_i^{(m)} + C_i \log\left[h_0(T_i|\mathbf{X}_i)\right] + w_i^{(m)} \log\left[S_0(T_i|\mathbf{X}_i)\right] \\ &= \sum_{i=1}^{n} C_i \log\left[w_i^{(m)} h_0(T_i|\mathbf{X}_i)\right] + w_i^{(m)} \log\left[S_0(T_i|\mathbf{X}_i)\right]. \end{aligned} \tag{2.11}$$

Note that when $C_i = 1$, $w_i^{(m)} = 1$ almost surely. Hence $C_i \log(w_i^{(m)}) = 0$ and $C_i w_i^{(m)} = C_i$ almost surely.

The "Maximisation" step maximises (2.10) and (2.11) to obtain updated estimates of $\Theta^{(m+1)} = \{\mathbf{b}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, h_b^{(m+1)}(T_i|B = 1)\}$. After that, we repeat the Expectation step with updated $\Theta^{(m+1)}$, and iterate until the estimated values for $\mathbf{b}$ and $\boldsymbol{\beta}$ converge. To deal with the nuisance function $h_b(T_i|B = 1)$, both the Breslow-type estimator and the product-limit estimator are suggested by Sy and Taylor (2000).

We first introduce the Breslow-type estimator. As discussed in Section 1.4.3, for a Cox proportional hazards model, Cox (1975) proposed to estimate $\boldsymbol{\beta}$ by partial likelihood (1.8). We will denote the estimated coefficients as $\hat{\boldsymbol{\beta}}$. Breslow (1972) suggested estimating $\boldsymbol{\beta}$ and the baseline cumulative hazard $H_b(t) = \int_0^t h_b(u)du$ in the maximum likelihood framework, using the joint likelihood

$$L(\beta, H_b) = \prod_{i=1}^{n} f(T_i)^{C_i} S(T_i)^{1-C_i} = \prod_{i=1}^{n} \Big\{ \exp(\boldsymbol{\beta}'\mathbf{X}_i) h_b(T_i) \Big\}^{C_i} \exp\Big(- \int_0^{T_i} \exp(\boldsymbol{\beta}'\mathbf{X}_i) h_b(u)du \Big).$$

It was shown that $L(\boldsymbol{\beta}, H_b)$ is maximised simultaneously at $\hat{\boldsymbol{\beta}}$ and

$$\hat{H}_b(t) = \sum_{i=1}^{n} \frac{I(T_i \leq t) C_i}{\sum_{j \in R(\tau_{(i)})} \exp(\hat{\boldsymbol{\beta}}'\mathbf{X}_i)}, \tag{2.12}$$

where $R(\tau_{(i)})$ is the index set for all individuals still at risk at the jth smallest uncensored survival time $\tau_{(i)}$. Details of the Breslow estimator can be found in Lin (2007).

To apply the Breslow estimator in the Expectation Maximisation algorithm context, we write

$$\begin{aligned} L_2(\boldsymbol{\beta}, H_b; w_i^{(m)}) &= \exp(\mathbb{E}[L_2]) \\ &= \prod_{i=1}^{n} \left[w_i^{(m)} h_b(T_i|B_i = 1) \exp(\boldsymbol{\beta}'\mathbf{X}_i) \right]^{C_i} S_b(T_i|B_i = 1)^{w_i^{(m)} \exp(\boldsymbol{\beta}'\mathbf{X}_i)}. \end{aligned}$$

It was shown that the Breslow nonparametric maximum likelihood estimator of $H_b(T_i|B_i = 1)$ given $\boldsymbol{\beta}$ is a slight modification to (2.12), whereas

$$\hat{H}_0(t|B = 1) = \sum_{i=1}^{n} \frac{I(T_i \leq t) C_i}{\sum_{j \in R(\tau_{(i)})} w_i^{(m)} \exp(\boldsymbol{\beta}'\mathbf{X}_i)}.$$

Using the above result in the Maximisation step, $\mathbb{E}[L_2]$ specifically, leads to a partial likelihood of $\boldsymbol{\beta}$

$$\prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}'\mathbf{X}_i)}{\sum_{j \in R(\tau_{(i)})} w_i^{(m)} \exp(\boldsymbol{\beta}'\mathbf{X}_i)} \right)^{C_i},$$

which can be maximised to find $\hat{\boldsymbol{\beta}}$.

Sy and Taylor (2000) suggested that in order to obtain a good estimate for $\mathbf{b}$ and $\boldsymbol{\beta}$, it is important for $\hat{S}_b(\tau_{(k)}|B = 1)$ to approach 0; recall that $\tau_{(k)}$ is the largest uncensored survival time. This constraint cannot be achieved using the Breslow-type estimator directly. However,

we can work around this issue by manually setting $\hat{S}_b(t|B=1) = 0$ for all $t > \tau_{(k)}$, as suggested and implemented in the R package "smcure" by Cai, Zou, Peng and Zhang (2012).

The baseline survival function can also be estimated using the product-limit estimator. As discussed in Kalbfleisch and Prentice (1980, p. 85), let

$$\alpha_j = S_b(\tau_{(j+1)})/S_b(\tau_{(j)}),$$

where $\tau_{(j)}$ is the jth smallest uncensored survival time, with covariates $\mathbf{X}_{(i)}$, and the α's are non-negative with $\alpha_0 = 1$. The baseline survival function can be computed as

$$S_b(t) = \prod_{j:\tau_{(j)} \leq t} \alpha_j,$$

with the probability mass only on distinct uncensored failure times, and $S_b(t^-_{(j)}) = S_b(t_{(j-1)})$. The corresponding hazard rate given $\boldsymbol{\beta}$ can be estimated as

$$\begin{aligned} h_0(\tau_{(j)}) &= \frac{\mathrm{d}F_0(\tau_{(j)})}{S_0(\tau_{(j)})} \\ &= \frac{S_0(\tau_{(j)}) - S_0(\tau_{(j+1)})}{S_0(\tau_{(j)})} \\ &= 1 - \frac{S_b(\tau_{(j+1)})^{\exp(\boldsymbol{\beta}'\mathbf{X})}}{S_b(\tau_{(j)})^{\exp(\boldsymbol{\beta}'\mathbf{X})}} \\ &= 1 - \alpha_j^{\exp(\boldsymbol{\beta}'\mathbf{X})}. \end{aligned}$$

Applying the resulting product-limit estimator in the Expectation Maximisation algorithm context yields

$$\begin{aligned} L_2(\boldsymbol{\beta}, \alpha_j; w_i^{(m)}) &= \exp(\mathbb{E}[L_2]) \\ &= \prod_{i=1}^{n} \left[w_i^{(m)} h_0(T_i|\mathbf{X}_i) \right]^{C_i} S_0(T_i|\mathbf{X}_i)^{w_i^{(m)}} \\ &= \prod_{i=1}^{n} \left(1 - \alpha_i^{\exp(\boldsymbol{\beta}'\mathbf{X}_i)} \right)^{C_i} S_b(T_i|B_i = 1)^{w_i^{(m)} \exp(\boldsymbol{\beta}'\mathbf{X}_i)} \\ &= \prod_{i=1}^{k} \left[\prod_{l \in D(\tau_{(i)})} (1 - \alpha_i^{\exp(\boldsymbol{\beta}'\mathbf{X}_l)}) \prod_{l \in (R(\tau_{(i)}) - D(\tau_{(i)}))} \alpha_i^{w_l^{(m)} \exp(\boldsymbol{\beta}'\mathbf{X}_l)} \right], \end{aligned}$$

where $D(\tau_{(i)})$ is the set of individuals experiencing the event of interest at time $\tau_{(i)}$.

Given $\boldsymbol{\beta}$, Sy and Taylor (2000) showed that the independent estimating equations for each α_j is

$$\sum_{l \in D(\tau_{(i)})} \frac{\exp(\boldsymbol{\beta}'\mathbf{X}_l)}{1-\alpha_i^{\exp(\boldsymbol{\beta}'\mathbf{X}_l)}} = \sum_{l \in R(\tau_{(i)})} w_l^{(m)} \exp(\boldsymbol{\beta}'\mathbf{X}_l),$$

and the maximum likelihood estimator for α_i when there are no ties at $\tau_{(i)}$ is

$$\hat{\alpha}_i = \left(1 - \frac{\exp(\boldsymbol{\beta}'\mathbf{X}_{(i)})}{\sum_{l \in R(\tau_{(i)})} w_l^{(m)} \exp(\boldsymbol{\beta}'\mathbf{X}_l)}\right)^{\exp(-\boldsymbol{\beta}'\mathbf{X}_{(i)})}.$$

We can then substitute $\hat{\alpha}_i$ into $L_2(\boldsymbol{\beta}, \alpha_j; w_i^{(m)})$ to obtain a nonparametric profile likelihood of $\boldsymbol{\beta}$ and obtain its maximum likelihood estimator. As suggested by Sy and Taylor (2000), when there are ties at $\tau_{(i)}$, the solution for α_i is not of closed form and the maximum likelihood estimators for $\boldsymbol{\beta}$ and α_i must be jointly obtained from $L_2(\boldsymbol{\beta}, \alpha_j; w_i^{(m)})$.

We note that such semiparametric Cox proportional hazards mixture cure models are always identifiable if a logistic regression is assumed for the incident model, according to Theorem 2.3.

2.2.5 Semiparametric Accelerated Failure Time Mixture Cure Model

As stated in Section 1.4.4, the accelerated failure time survival model can be either parametric or semiparametric. Unlike the generalised gamma accelerated failure time mixture cure model as discussed in Section 2.2.3, Li and Taylor (2002) utilised the accelerated failure time framework from a semiparametric point of view. Specifically, an accelerated failure time model as defined by (1.4.4) is semiparametric when the distribution for ϵ is assumed to be nonparametric. In the mixture cure model context, Li and Taylor (2002) assumed the survival function for the susceptibles, $S_0(t)$, follows a semiparametric accelerated failure time model with a slight modification, such that

$$\log T^*|_{B=1} = \boldsymbol{\beta}'\mathbf{X} + \epsilon,$$

where the error term ϵ has a density function f_ϵ and a distribution function F_ϵ. Therefore, we write

$$P(T^* < t|B=1) = F_\epsilon(\log T^* - \boldsymbol{\beta}'\mathbf{X}).$$

Assuming as usual a logistic regression (2.2) for the incident model, given an observed survival dataset $\mathcal{D} = \{T, C, \mathbf{X}\}$, the contribution of an uncensored observation i to the complete

likelihood is

$$p_i f_\epsilon(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i)\Big|\frac{\mathrm{d}\epsilon}{\mathrm{d}T}\Big| = \frac{p_i f_\epsilon(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i)}{T_i},$$

again with $p_i = p(\mathbf{X}_i)$. On the other hand, the contribution of a censored observation j to the complete likelihood is

$$1 - p_j + p_j S_\epsilon(\log T_j^* - \boldsymbol{\beta}'\mathbf{X}_j).$$

Therefore, the complete likelihood given observed data in this case is

$$L_{obs}(\mathbf{b}, \boldsymbol{\beta}, F_\epsilon; \mathcal{D}) = \prod_{i=1}^{n} \Big\{ \frac{p_i f_\epsilon(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i)}{T_i} \Big\}^{C_i} \Big\{ 1 - p_i + p_i\big[1 - F_\epsilon(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i)\big] \Big\}^{1-C_i},$$

and the corresponding log-likelihood is

$$\begin{aligned} l_{obs}(\mathbf{b}, \boldsymbol{\beta}, F_\epsilon; \mathcal{D}) &= \sum_{i=1}^{n} C_i \log\Big\{ p_i f_\epsilon(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i) \Big\} - \sum_{i=1}^{n} C_i \log T_i^* + \\ &\qquad \sum_{i=1}^{n} (1 - C_i) \log\Big\{ 1 - p_i + p_i\big[1 - F_\epsilon(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i)\big] \Big\} \\ &\propto \sum_{i=1}^{n} C_i \log\Big\{ p_i f_\epsilon(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i) \Big\} + \\ &\qquad \sum_{i=1}^{n} (1 - C_i) \log\Big\{ 1 - p_i + p_i\big[1 - F_\epsilon(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i)\big] \Big\}, \end{aligned}$$

since the term $\sum_{i=1}^{n} C_i \log(T_i)$ is not dependent on the parameters $(\mathbf{b}, \boldsymbol{\beta}, F_\epsilon)$ of interest.

To estimate the regression coefficients, Li and Taylor (2002) suggested using the Expectation Maximisation algorithm. To implement it, we need to use the latent variable B, the susceptible indicator. This variable is unobserved but completes the data when combined with the observed variables, where the complete data is $\mathcal{D}_C := \{T, \mathbf{X}, C, B\}$. The full likelihood, disregarding the constant factors, given the complete data can then be written as

$$\begin{aligned} L_C(\mathbf{b}, \boldsymbol{\beta}, F_\epsilon; t, \mathcal{D}_C) &= \prod_{i=1}^{n} \Big\{ (1 - p_i) \Big\}^{C_i(1-B_i)} \Big\{ \frac{p_i f_\epsilon(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i)}{T_i} \Big\}^{C_i B_i} \Big\{ (1 - p_i) \Big\}^{(1-B_i)(1-C_i)} \times \\ &\qquad \Big\{ p_i\big[1 - F_\epsilon(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i)\big] \Big\}^{B_i(1-C_i)} \\ &= \prod_{i=1}^{n} p_i^{B_i} (1 - p_i)^{(1-B_i)} \prod_{i=1}^{n} \Big(\frac{f_\epsilon(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i)}{T_i} \Big)^{C_i B_i} \big[1 - F_\epsilon(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i)\big]^{B_i(1-C_i)} \\ &= L_1(\mathbf{b}|Y) \times L_2(\boldsymbol{\beta}, F_\epsilon|Y). \end{aligned}$$

Taking log transforms, we have

$$l_1(\mathbf{b}|Y) = \sum_{i=1}^{n} \Big\{ B_i \log p_i + (1 - B_i) \log(1 - p_i) \Big\}, \tag{2.13}$$

$$l_2(\beta, F_\epsilon|Y) \propto \sum_{i=1}^{n} \Big\{ C_i B_i \log \left(f_\epsilon(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i) \right) + B_i(1 - C_i) \log \left[1 - F_\epsilon(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i) \right] \Big\}. \tag{2.14}$$

As mentioned in Section 2.2.4, the Expectation step of the Expectation Maximisation algorithm takes expectation of the two components of $l_C(\mathbf{b}, \boldsymbol{\beta}, F_\epsilon; \mathcal{D}_C)$ with respect to the unobserved B given the current parameter values (at the $(m+1)$-th iteration), $\Theta^{(m)} = \{\mathbf{b}^{(m)}, \boldsymbol{\beta}^{(m)}, F_\epsilon^{(m)}\}$ and the observed data, $O = \{\text{observed } B, t, c, \mathbf{X}\}$. For censored cases, both $l_1(\mathbf{b}|B, \mathbf{X})$ and $l_2(\boldsymbol{\beta}, F_\epsilon|B, \mathbf{X})$ are linear functions of B_i, therefore, the conditional expectation of B_i will be needed to complete the Expectation step. The computation is very similar to that discussed in Section 2.2.4, the result is also very similar to Equation (2.9)

$$w_i^{(m)} = \mathbb{E}[B_i|\Theta^{(m)}, O] = C_i + (1 - C_i) \frac{p_i \left[1 - F_\epsilon(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i) \right]}{1 - p_i + p_i \left[1 - F_\epsilon(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i) \right]} \Big|_{\Theta^{(m)}}. \tag{2.15}$$

Replacing the B_i's with $w_i^{(m)}$ defined by (2.15), the expectations of (2.13) and (2.14) are

$$\mathbb{E}[l_1(\mathbf{b}|B, \mathbf{X})] = \sum_{i=1}^{n} \Big\{ w_i^{(m)} \log p_i + (1 - w_i^{(m)}) \log(1 - p_i) \Big\},$$

$$\mathbb{E}[l_2(\boldsymbol{\beta}|B, \mathbf{X})] = \sum_{i=1}^{n} \Big\{ C_i w_i^{(m)} \log \left(f_\epsilon(\log T_i^* - \beta'\mathbf{X}_i) \right) + w_i^{(m)}(1 - C_i) \log \left[1 - F_\epsilon(\log T_i^* - \beta'\mathbf{X}_i) \right] \Big\}.$$

The Maximisation step starts by first updating $\mathbf{b}^{(m)}$ with the new $\mathbf{b}^{(m+1)}$, which can be obtained by maximising $\mathbb{E}[l_1]$ given $w^{(m)}$.

To find the estimators for $\boldsymbol{\beta}$ in the maximum likelihood framework, we find the estimating equation by setting

$$\frac{\mathrm{d}\mathbb{E}[l_2(\boldsymbol{\beta}|B, \mathbf{X})]}{\mathrm{d}\beta} = 0,$$

which translates into

$$\sum_{i=1}^{n} \mathbf{X}_i \Big\{ - C_i w_i^{(m)} \frac{f_\epsilon'}{f_\epsilon} \big(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i \big) + w_i^{(m)}(1 - C_i) \frac{f_\epsilon}{1 - F_\epsilon} \big(\log T_i^* - \boldsymbol{\beta}'\mathbf{X}_i \big) \Big\} = 0.$$

While the equation is difficult to compute due to the presence of $f_\epsilon(t)$ and $f_\epsilon'(t)$, Ritov (1990) suggested proceeding with an M-estimator, which can be achieved by replacing $-f_\epsilon'/f_\epsilon$ with

any reasonable score function s, so the estimating equation becomes

$$\sum_{i=1}^{n} \mathbf{X}_i \Big\{ C_i w_i^{(m)} s\big(\log T_i^* - \boldsymbol{\beta}' \mathbf{X}_i \big) + w_i^{(m)} (1 - C_i) \frac{\int_{\log T_i^* - \boldsymbol{\beta}' \mathbf{X}_i}^{\infty} s(u) dF_\epsilon(u)}{1 - F_\epsilon(\log T_i^* - \boldsymbol{\beta}' \mathbf{X}_i)} \Big\} = 0.$$

In the examples given by Li and Taylor (2002), they considered two choices of $s(u)$. One is the identity

$$s(u) = u,$$

and the other is the bounded function

$$s(u) = \begin{cases} -3 & u < -3 \\ u & |u| \leq 3 \\ 3 & u > 3. \end{cases}$$

Recall that f_ϵ is known only up to a shift since the accelerated failure time model does not include an intercept term. Hence, as noted by Ritov (1990), the estimating equation does not have a mean of zero when we plug in the correct $\boldsymbol{\beta}$ with a misspecified intercept. To remedy this issue, we centre the covariates $\mathbf{X}$ so that the estimating equation becomes

$$\sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}}) \Big\{ C_i w_i^{(m)} s\big(\log T_i^* - \boldsymbol{\beta}' \mathbf{X}_i \big) + w_i^{(m)} (1 - C_i) \frac{\int_{\log T_i^* - \boldsymbol{\beta}' \mathbf{X}_i}^{\infty} s(u) dF_\epsilon(u)}{1 - F_\epsilon(\log T_i^* - \boldsymbol{\beta}' \mathbf{X}_i)} \Big\} = 0.$$

where $\bar{\mathbf{X}} = \sum_{i=1}^{n} \mathbf{X}_i / n$.

To handle the unknown F_ϵ, the distribution function of the residuals, we first consider the simple case where $p = 1$ so that there are no immunes in the data. In that case, Buckley and James (1979) suggested replacing F_ϵ with the generalised maximum likelihood estimator, which is the nonparametric Kaplan-Meier estimator in this case. Ritov (1990) proved that the estimator for $\boldsymbol{\beta}$ suggested by Buckley and James (1979) is consistent and asymptotically normal. When a cured proportion is considered, the nonparametric estimator can be generalised, as suggested by Taylor (1995). Let $\tau_{(1)}, \cdots, \tau_{(k)}$ denote the distinct observed failure times, with d_j tied failures at time $\tau_{(j)}$. Let H_j denote the set of $n_{c,j}$ censored observations with censored times in the interval

$[\tau_{(j)}, \tau_{(j+1)})$. Taylor (1995) suggests

$$\begin{aligned} f_0(\tau_{(j)}) &= h_0(\tau_{(j)}) \prod_{l=1}^{j-1} \left(1 - h_0(\tau_{(l)})\right) \\ S_0(t) &= \prod_{j:\tau_{(j)}<t} \left(1 - h_0(\tau_{(l)})\right) \\ S_0(\tau_{(j)}) &= \prod_{l=1}^{j-1} \left(1 - h_0(\tau_{(l)})\right) \\ S_0(\tau_{(j)} + 0^+) &= \prod_{l=1}^{j} \left(1 - h_0(\tau_{(l)})\right), \end{aligned}$$

where again, $h(t)$ denotes the hazard function defined in (1.4.3). Let w_l be the conditional probability of a censored observation l being an immune, so that for $l \in H_j$,

$$w_l = \frac{p_l S_0(\tau_{(j)} + 0^+)}{1 - p_l + p_l S_0(\tau_{(j)} + 0^+)},$$

where $p_l = p(\mathbf{X}_l)$. Consequently, we write the likelihood function for the complete data as

$$\prod_{j=1}^{k} \left\{ \left[f_0(\tau_{(j)}) \right]^{d_j} \prod_{l \in H_j} \left[S_0(\tau_{(j)} + 0^+) \right]^{w_l} \right\},$$

which is maximised by

$$h_0(\tau_{(j)}) = \frac{d_j}{\sum_{r=j}^{k} (d_r + \sum_{l \in H_r} w_l)}.$$

Clearly, the estimator for $h(\tau_j)$ reduces to the Kaplan-Meier estimator when there are no cured observations, so $w_l = 1$. The distribution function for the residuals can then be obtained since $S_0(T_i) = S_\epsilon(\log T_i^* - \boldsymbol{\beta}' \mathbf{X}_i)$. Using this result, the Expectation Maximisation algorithm can then be used to provide estimators for $\mathbf{b}$ and $\boldsymbol{\beta}$.

In addition, the methodology was later studied by several authors. For example, with similar model setups, Zhang and Peng (2007) modified the estimation method using a rank-based estimation method to estimate $\boldsymbol{\beta}$. In contrast, Lu (2010) maximised a kernel-smoothed conditional profile likelihood in the Maximisation step for the Expectation Maximisation algorithm and showed that the estimator is consistent and asymptotically normal. Again, we note that the semiparametric accelerated failure time mixture cure models are always identifiable if a logistic regression is assumed for the incident model, according to Theorem 2.3.

2.3 Non-Mixture Cure Models

Non-mixture cure models were introduced by Yakovlev, Bardou, Asselain, Fourquet, Hoang, Rochefordiere and Tsodikov (1993), and they originated from a biological motivation. Specifically, assume that after treatment, a patient is left with $N \sim Poisson(\theta)$ cancer cells that may grow to a detectable disease sometime later. These N cancer cells are referred to as competing risks. If one of the competing risks fails, i.e. one of the cancer cells grows to a detectable disease, the individual will inevitably experience the event of interest and belong to the susceptible group. On the other hand, if the patient has $N = 0$ cancer cell left, he/she becomes immune to the event of interest. Again we let T_i^* be the actual survival time for the i-th individual that is in the susceptible group, but in this section, we define it as

$$T_i^* = \min\{T_{i,1}^*, ..., T_{i,N_i}^*\}, \tag{2.16}$$

where $T_{i,j}^*$'s are the latent survival times of the j-th competing risk for the i-th individual. They are assumed to be independent and identically distributed with a proper cumulative distribution function $F^*(t) = 1 - S^*(t)$. The overall survival function for the population is then, as noted by Peng and Taylor (2014),

$$\begin{aligned}
S(t) &= P(N = 0) + P(T_{i,1}^* \geq t, \cdots, T_{i,N_i}^* \geq t; N \geq 1) \\
&= \exp(-\theta) + \sum_{k=1}^{\infty} S^*(t)^k \frac{\theta^k}{k!} \exp(-\theta) \\
&= \sum_{k=0}^{\infty} S^*(t)^k \frac{\theta^k}{k!} \exp(-\theta) \\
&= \sum_{k=0}^{\infty} \frac{\big[\theta S^*(t)\big]^k}{k!} \exp\big(-\theta S^*(t)\big) \exp\big(-\theta + \theta S^*(t)\big) \\
&= \exp\big(-\theta F^*(t)\big).
\end{aligned}$$

This is an improper survival function since $S(\infty)$ does not equal 0. The cure fraction in this model is defined as

$$1 - p = P(N = 0) = \exp(-\theta), \tag{2.17}$$

in other words, if a patient does not have any "competing" cancer cells left in the body, they are completely cured by the treatment. When we assume the cure fraction is dependent on the

covariates $\mathbf{X}$, unlike the mixture cure models, we simply relate the parameter θ to the covariates, for example,

$$\theta(\mathbf{X}) = \exp(\mathbf{b}'\mathbf{X}). \tag{2.18}$$

If we implement (2.18), the survival function is

$$S(t) = \exp\left(-\exp(\mathbf{b}'\mathbf{X})F^*(t)\right).$$

Definition 2.3 (Non-mixture Cure Model)**.** Given survival data $\mathcal{D} = \{T, C, \mathbf{X}\}$ with T^* defined by (2.16). Let $S^*(t) = 1 - F^*(t)$ be a proper survival function. Assume the number of competing risks for each individual, N, to follow a Poisson distribution with mean $\theta(\mathbf{X}) \geq 0$. The survival function of a non-mixture cure model can be expressed as

$$S(t) = \exp\big(-\theta(\mathbf{X})F^*(t)\big), t > 0,$$

where the relationship between F^* and the covariates depends on the model of choice.

Aside from the biological motivation, Chen, Ibrahim and Sinha (1999) emphasised that the non-mixture model is suitable for any failure time data, as long as the data has a cured fraction and can be thought of as generated by an unknown number N of latent competing risks. In addition, Tsodikov (1998) proposed a different motivation for a non-mixture cure model, which does not directly relate to a biological background. Let $H(t)$ be the corresponding cumulative hazard for the overall population. By Equations (1.6) and (1.7), the overall population survival function is

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

Note that when a cured fraction is considered, $S(t)$ is improper, and consequently $H(t)$ is bounded such that

$$H(t) \leq \theta.$$

Equivalently,

$$\lim_{t\to\infty} H(t) = \theta.$$

A convenient way to adjust for this is by considering $H(t) = \theta F^*(t)$, where $F^*(t)$ is the proper distribution function of a non-negative random variable. This will lead to the model structure

in Definition 2.3.

For a non-mixture cure model defined by Definition 2.3, the corresponding density is computed as

$$f(t) = \theta(\mathbf{X}) f^*(t) \exp\big(-\theta(\mathbf{X}) F^*(t)\big),$$

and the hazard function is

$$h(t) = \frac{f(t)}{S(t)} = \theta(\mathbf{X}) f^*(t) = \exp(\mathbf{b}'\mathbf{X}) f^*(t), \tag{2.19}$$

when (2.18) is used to model the relationship between θ and the covariates $\mathbf{X}$. From (2.19), it is clear that the non-mixture cure model has a proportional hazards structure, where the baseline hazard is $f^*(t)$. The hazard ratio is $\exp(\mathbf{b}'\mathbf{X})$ and it does not depend on the survival time T^*. For this reason, a non-mixture cure model can also be referred to as a proportional hazards cure model.

As noted by Peng and Taylor (2014), the drawback of such a non-mixture model is that the effect of covariate $\mathbf{X}$ on the susceptible survival distribution cannot be interpreted easily, because it does not separate the immune and the susceptible individuals like a mixture cure model.

The survival function for the susceptibles can be calculated as (using Bayes' theorem and the law of total probability)

$$\begin{aligned} S_0(t) = P(T_i^* > t | N \geq 1) &= \frac{P(T_i^* > t, N \geq 1)}{P(N \geq 1)} \\ &= \frac{P(T_i^* > t) - P(T_i^* > t, N = 0)}{P(N \geq 1)} \\ &= \frac{\exp\big(-\theta F^*(t)\big) - \exp\big(-\theta\big)}{1 - \exp\big(-\theta\big)}. \end{aligned}$$

Since $F^*(t)$ is a proper distribution function and hence right continuous and non-decreasing, it is clear that $S_0(t)$ is right continuous and non-increasing with $S_0(0) = 1$ and $S_0(\infty) = 0$. Therefore, $S_0(t)$ is a proper survival function. Its corresponding density is

$$f_0(t) = \frac{\exp\big(-\theta F^*(t)\big)}{1 - \exp(-\theta)} \theta f^*(t),$$

and the hazard function for the susceptibles is

$$h_0(t) = \frac{f_0(t)}{S_0(t)} = \left(\frac{\exp\big(-\theta F^*(t)\big)}{\exp\big(-\theta F^*(t)\big) - \exp(-\theta)}\right)\theta f^*(t) = \left(\frac{1}{P(T_i^* < \infty | T_i^* > t)}\right)h(t),$$

where

$$P(T_i^* < \infty | T_i^* > t) = \frac{P(T_i^* < \infty, T_i^* > t)}{P(T_i^* > t)} = \frac{P(T_i^* > t, N \geq 1)}{P(T_i^* > t)} = \frac{\exp\big(-\theta F^*(t)\big) - \exp(-\theta)}{\exp\big(-\theta F^*(t)\big)}.$$

Thus the hazard function for the susceptibles is not proportional because the hazard ratio always depends on T^*.

Chen et al. (1997) pointed out that there is a connection between the non-mixture cure models and the mixture cure models. Specifically, using (2.17), a non-mixture model will take the form

$$\begin{aligned} S(t) &= \exp\big(-\theta F^*(t)\big) \\ &= \exp(-\theta) + \exp\big(-\theta F^*(t)\big) - \exp(-\theta) \\ &= \exp(-\theta) + \big(1 - \exp(-\theta)\big)\frac{\exp\big(-\theta F^*(t)\big) - \exp(-\theta)}{1 - \exp(-\theta)} \\ &= (1-p) + pS_0(t), \end{aligned}$$

which is in the form of a mixture cure model defined by Definition 2.1. Hence, every non-mixture cure model can essentially be written as a mixture cure model. However, the fundamental differences between them remain: non-mixture models model the entire population as a proportional hazards model, and the hazards for susceptibles are not proportional; semiparametric Cox proportional hazards mixture cure models can model only the susceptible group as a proportional hazards model, and the hazards for the whole population are not proportional.

In terms of the identifiability issues for non-mixture cure models, we again refer to Li, Taylor and Sy (2001), where their results are summarised by Definition 2.4 and Theorem 2.4 below.

Definition 2.4 (Definition 2 in Li, Taylor and Sy (2001))**.** Following Definition 2.3, let $\mathcal{F}^*$ be a class of proper lifetime distributions on $[0, \infty)$, $\Theta^* = \{\theta(X) : \theta(X) \neq 0 \text{ or } \infty, \text{ for all } X \in \mathcal{X}\}$ (assuming $\mathcal{X} = [a_0, a_1]$). The class of non-mixture cure models, defined by

$$\mathcal{S} = \{S(t;X) : S(t;X) = \exp(-\theta(X)F^*(t), X \in \mathcal{X}, \theta(\cdot) \in \Theta^*, F^*(\cdot) \in \mathcal{F}^*\},$$

is identifiable if for $S(t;X)$ and $S_{(1)}(t;X)$, two members in $\mathcal{S}$, $S(t;X)$ is identical to $S_{(1)}(t;X)$ if and only if $\theta(X)$ is identical to $\theta_{(1)}(X)$ for $X \in \mathcal{X}$ and $F^*(t)$ is identical to $F^*_{(1)}(t)$ for almost all $t > 0$.

Theorem 2.4 (Theorems 4 and 5 in Li, Taylor and Sy (2001))**.** *For a non-mixture cure model as defined by Definition 2.3 and X being a continuous one-dimensional covariate,*

1. *when $F^*(t)$ and $\theta(X)$ are both unspecified, the model is not identifiable for X in the design space;*
2. *when $F^*(t)$ is unspecified, and $\theta(X) = \exp(a + \boldsymbol{b}'X)$, the model is not identifiable unless the intercept term $a = 0$;*
3. *when $F^*(t)$ is specified in a parametric form determined by a finite number of parameters, the model is identifiable on $X \in \mathcal{X}$, regardless of whether $\theta(X)$ is unspecified or $\theta(X) = \exp(a + \boldsymbol{b}'X)$.*

In the following subsections, we list both the frequentist and the Bayesian methods used in literature to estimate the parameters in a non-mixture cure model.

2.3.1 Tsodikov (1998) Maximum Likelihood Approach

Given a set of survival data $\mathcal{D} = \{T, C, \mathbf{X}\}$, recall that a non-mixture cure model as defined in Definition 2.3, where the relationship between θ and the observed covariates $\mathbf{X}$ is modelled by (2.18), has a proportional hazards structure as shown by (2.19). Hence, it is natural for us to first define some baseline θ_b and parameterise the ratio

$$\gamma_r = \theta(\mathbf{X})/\theta_b.$$

Following a similar approach proposed by Cox (1975), Tsodikov (1998) denoted the untied observed uncensored survival times as $\tau_{(1)} < \tau_{(2)} < \cdots < \tau_{(k)}$ in an increasing order. Suppose that $n_{c,i}$ censored observations are observed in the interval $(\tau_{(i)}, \tau_{(i+1)})$, for $i = 1, ..., k$, where $\tau_{(k+1)}$ is defined to be infinite. Denote their censoring times as $\tau_{(i),1}, ..., \tau_{(i),n_{c,i}}$ respectively, and each corresponds with a $\theta^c_{i,j} = \theta(\mathbf{X}_{(i),j})$ for $j = 1, ..., n_{c,i}$. Let θ^d_i denote the value of $\theta(\mathbf{X}_{(i)})$ for the individual who fails at time $\tau_{(i)}$, and Θ_i denote the sum of θ's over all the individuals in the interval $[\tau_{(i)}, \tau_{(i+1)})$ (i.e. $\Theta_i = \theta^d_i + \sum_{j=1}^{n_{c,i}} \theta^c_{i,j}$), where we assume any censored observations tied with failures are moved a small amount to the right. The complete likelihood given data $\mathcal{D}$ is

then

$$L = \prod_{i=1}^{k} \theta_i^d f^*(\tau_{(i)}) \exp\left(-\theta_i^d F^*(\tau_{(i)})\right) \prod_{j=1}^{n_{c,i}} \exp\left(-\theta_{i,j}^c F^*(\tau_{i,j})\right),$$

where $\tau_{i,j}$ denotes the j-th observation among all $n_{c,i}$ censored observations in the interval $[\tau_{(i)}, \tau_{(i+1)})$.

A partial likelihood L_p contains sufficient information to estimate the parameters involved in γ_r, where

$$L_p = \frac{\prod_{i=1}^{k} \theta_i^d}{S_{1,k} S_{2,k} \cdots S_{k,k}}, \tag{2.20}$$

with $S_{a,b} = \sum_{j=a}^{b} \Theta_j$, assuming the convention that $\sum_a^b = 0$ and $\prod_a^b = 1$ for all $a > b$. Tsodikov (1998) claimed that the likelihood (2.20) provides consistent estimators of the ratios γ_r, but it should be noted that the partial likelihood is useless if estimation of the baseline θ_b is of interest; in fact both θ_b and $\theta(\mathbf{X})$ are often of interest when cure is a possibility.

There is not enough information in the partial likelihood (2.20) to estimate $\mathbf{b}$ and hence $\theta(\mathbf{X})$. Tsodikov (1998) further suggested using a marginal likelihood. Recall in Section 1.3 we assumed that the times to censoring events are independent and identically distributed with cumulative distribution function G, and they are independent of the times to failure. The likelihood of the data is given by

$$L = \prod_{i=1}^{k} \theta_i^d f^*(\tau_{(i)}) \exp(-\theta_i^d F^*(\tau_{(i)})) \left(1 - G(\tau_{(i)})\right) \prod_{j=1}^{n_{c,i}} \exp\left(-\theta_{i,j}^c F^*(\tau_{i,j})\right) \Delta\left(1 - G(\tau_{i,j})\right), \tag{2.21}$$

where $\Delta(1 - G(t)) = P(\text{censored at } t)$. By integrating (2.21) over the set $\{\tau_{(1)} < \tau_{(2)} < \cdots < \tau_{(k)}$ and $\tau_{(i)} < \tau_{i,j} < \tau_{(i+1)}\}$, we obtain the probability of observing a given ordering of uncensored times and $n_{c,i}$ censored observations between successive failures, and it generally depends on the distributions G and F. As suggested by Tsodikov (1998), this probability might need to be considered to find the rank marginal likelihood. Tsodikov (1998) also showed the marginalisation succeeds in eliminating the nuisance parameters if censoring is of Type I (the study ends at a predetermined time t_c, where the remaining individuals are then right-censored). In such a case with Type I censoring, G will be a single-step function, such that $G(\tau_{(i)}) = 1$, $i = 1, ...k$; $\Delta G(t_c) = 1$; $\tau_{k,j} = t_c$, and we may write the marginal likelihood L_m as

$$L_m = \int_0^{t_c} \int_{\tau_{(1)}}^{t_c} \cdots \int_{\tau_{(k-1)}}^{t_c} \prod_{i=1}^{k} \theta_i^d f^*(\tau_{(i)}) e^{-\theta_i^d F^*(\tau_{(i)})} e^{-(\Theta_k - \theta_k^d) F^*(t_c)} \mathrm{d}\tau_{(k)} \cdots \mathrm{d}\tau_{(2)} \mathrm{d}\tau_{(1)} = L_p L_c L_d,$$

where

$$L_c = e^{-\tilde{\Theta}_k F(t_c)}, \quad L_d = \sum_{i=0}^{k} (-1)^i \exp\{-\tilde{S}_{k-i+1,k} F^*(t_c)\} \frac{P_{1,k}}{P_{1,k-i} P^{k-i+1,k}},$$

and

$$\tilde{S}_{a,b} = \sum_{j=a}^{b} \theta_j^d, \quad \tilde{\Theta}_k = \Theta_k - \theta_k^d,$$

$$P_{a,b} = \tilde{S}_{a,b} \tilde{S}_{a+1,b} \cdots \tilde{S}_{b,b}, \quad P^{a,b} = \tilde{S}_{a,b} \tilde{S}_{a,b-1} \cdots \tilde{S}_{a,a}.$$

Tsodikov interpreted the components L_p as the partial likelihood as shown in (2.20); L_c as the probability of observing certain censored individuals at the end of the study; and L_d as specific to failures.

When F^* is completely unknown, the marginal likelihood is not as efficient as maximising the profile likelihood, which we denote as l_{pr}. Tsodikov (1998) introduced the profile likelihood as the full likelihood maximised with respect to F, with the nonparametric maximum likelihood estimator used to estimate F. Tsodikov (1998) noted that since the nonparametric maximum likelihood estimator does not imply untied observations, we will assume that there are $d_i \geq 1$ failures at an observed failure time $\tau_{(i)}$, the corresponding θ's are denoted as $\theta_{i,j}^d$, for $i = 1, ...k$, $j = 1, ..., d_i$. Tsodikov (1998) presented the full log-likelihood as

$$l_{pr} = \sum_{i=1}^{k} \Big\{ \sum_{j=1}^{d_i} \log \left[e^{-\theta_{i,j}^d F^*(\tau_{(i)} - 0^+)} - e^{-\theta_{i,j}^d F^*(\tau_{(i)})} \right] - \sum_{j=1}^{n_{c,i}} \theta_{i,j}^c F^*(\tau_{(i)}) \Big\}. \tag{2.22}$$

Tsodikov (1998) further suggested that the likelihood (2.22) can be maximised with respect to F^* by first estimating F^* by a nonparametric maximum likelihood estimator. That is, we replace F^* by a step-function taking values at the failure times so that

$$F^*(\tau_{(i)-0^+}) = F^*(\tau_{(i-1)}), \quad \text{for } i = 0, ..., k; \quad F^*(\tau_{(0)}) = 0; \quad F^*(\tau_{(n)}) = 1.$$

2.3.2 Chen, Ibrahim, Sinha's (1999) Bayesian Approach

As an alternative to the frequentist approach, Chen, Ibrahim and Sinha (1999) utilised the Bayesian method to estimate the parameters of interest. Let the observed survival data of n individuals be $\mathcal{D} = \{T, C, \mathbf{X}\}$. The complete data also includes the unobserved latent variable

N, which is the number of competing risks, for each individual, so it is of the form

$$\mathcal{D}_C = \{T, C, \mathbf{X}, N\}.$$

Recall that $N_i = 0$ indicates that the ith individual is a long-term survivor.

Chen et al. (1999) assumed a parametric distribution for $F^*(t)$ specified by the vector of parameters Θ, whose dimension is finite. For example, if a Weibull distribution is specified, then $\Theta = (\alpha, \lambda)'$, where α is the shape parameter and λ is the scale parameter. Hence for a censored observation indexed as i, the contribution to the overall likelihood is

$$S(T_i|\Theta)^{N_i} \times \frac{\exp(-\theta_i)\theta_i^{N_i}}{N_i!},$$

since none of the N_i competing risks fails up to the censored time. For an uncensored observation j, the contribution to the likelihood is

$$S(t_j|\Theta)^{N_j-1} \times \binom{N_j}{1} \times f(t_j|\Theta) \times \frac{\exp(-\theta_j)\theta_j^{N_j}}{N_j!},$$

as the individual experiences the event when the first failure among the N_j competing risks is observed. Combining the results, the complete-data likelihood function of the data given parameters (Θ, θ) can be written as

$$L(\mathbf{b}, \Theta|\mathcal{D}_C) = \Big(\prod_{i=1}^{n} S(T_i|\Theta)^{N_i - C_i} \big(N_i f(T_i|\Theta)\big)^{C_i}\Big) \times \exp\Big(\sum_{i=1}^{n} \big((\mathbf{b}'\mathbf{X}_i)N_i - \log N_i! - \exp(\mathbf{b}'\mathbf{X}_i)\big)\Big). \tag{2.23}$$

To construct the likelihood function (2.23), Chen et al. (1999) suggested assuming a Weibull distribution for $F^*(t)$, so

$$S^*(t) = \exp(-\lambda t)^{\alpha},$$
$$f^*(t) = \alpha\lambda(\lambda t)^{\alpha-1} \exp\{-(\lambda t)^{\alpha}\}.$$

To estimate the parameters using the Bayesian method, we need to specify a prior distribution, and then use Bayes' Rule to compute the posterior distribution. For the prior, Chen et al. (1999) considered a joint non-informative prior for $(\mathbf{b}, \Theta = \{\alpha, \lambda\})$, where the parameters are

independent of each other and $\pi(\mathbf{b}) \propto 1$ is a uniform improper prior. Consequently,

$$\pi(\mathbf{b}, \Theta) \propto \pi(\Theta).$$

With a Weibull distribution specified, while not necessary, we can further assume a gamma distribution with hyperparameters prior ϱ_0 and ν_0 as a prior for the parameter α, where

$$\pi(\Theta) = \pi(\alpha|\varrho_0, \nu_0)\pi(\lambda),$$

and

$$\pi(\alpha|\varrho_0, \nu_0) \propto \alpha^{\varrho_0 - 1} \exp(-\nu_0 \alpha).$$

Using Bayes' Rule, the posterior distribution for $(\mathbf{b}, \Theta)$ based on the observed data $\mathcal{D}$ can be expressed as

$$p(\mathbf{b}, \Theta|\mathcal{D}) \propto \Big(\sum_N L(\mathbf{b}, \Theta|\mathcal{D}_C) \Big) \pi(\alpha|\varrho_0, \nu_0)\pi(\lambda), \tag{2.24}$$

where the sum in (2.24) extends over all possible values of N.

Informative priors can be constructed from previous similar studies that measure the same response variable and covariates as the current study. The data from that similar previous study is referred to as the historical data. Let the observed historical data of n_0 observations be

$$\mathcal{D}_0 = \{T_0, \mathbf{X}_0, C_0\},$$

and the complete data be $\mathcal{D}_{C,0} = \{T_0, \mathbf{X}_0, C_0, N_0\}$, where N_0 is again unobserved. An informative prior can then be computed as

$$\pi(\mathbf{b}, \Theta|\mathcal{D}_0, a_0) \propto \left[\sum_{N_0} L(\mathbf{b}, \Theta|\mathcal{D}_{C,0}) \right]^{a_0} \pi_0(\mathbf{b}, \Theta),$$

where $\pi_0(\mathbf{b}, \Theta)$ is the initial prior distribution from the previous study, and $L(b_0, b_1, \Theta|\mathcal{D}_{C,0})$ is the complete-data likelihood. The extra parameter a_0 controls the "level" for the informative prior, where a small a_0 indicates a less informative prior, and when $a_0 = 1$, the informative prior becomes the posterior from the previous study. It is reasonable to take $0 \leq a_0 \leq 1$.

In practice, Chen et al. (1999) suggested using a non-informative prior for $\pi_0(\mathbf{b}, \Theta)$, such that

$$\pi_0(\mathbf{b}, \Theta) \propto \pi_0(\Theta),$$

where $\pi_0(\mathbf{b}) \propto 1$ and $\Theta = (\alpha, \lambda)$, if a Weilbull distribution is specified for F^*. They further suggested taking a gamma prior with some arbitrary small shape and scale parameters for α; an independent normal prior with mean 0 and variance c_0 for λ; and, a beta prior with arbitrary parameter values for a_0. Hence, the joint informative prior for $(\mathbf{b}, \Theta, a_0)$ is of the form

$$\pi(\mathbf{b}, \Theta, a_0 | \mathcal{D}_0) \propto \left[\sum_{N_0} L(\mathbf{b}, \Theta | \mathcal{D}_{C,0}) \right]^{a_0} \pi_0(\mathbf{b}, \Theta) a_0^{c_0 - 1} (1 - a_0)^{\lambda_0 - 1}, \tag{2.25}$$

where c_0 and λ_0 are the hyperparameters for the beta prior.

The joint posterior given the prior (2.25) based on the current observed data $\mathcal{D}$ is then

$$\pi(\mathbf{b}, \Theta, a_0 | \mathcal{D}_0, \mathcal{D}) \propto \left(\sum_{N} L(\mathbf{b}, \Theta | \mathcal{D}_C) \right) \left[\sum_{N_0} L(\mathbf{b}, \Theta | \mathcal{D}_{C,0}) \right]^{a_0} \pi_0(\mathbf{b}, \Theta) a_0^{c_0 - 1} (1 - a_0)^{\lambda_0 - 1}.$$

The Markov chain Monte Carlo algorithm can then be used to sample from the joint posterior density and obtain the estimators.

Chapter 3

Problems of Insufficient Follow-up

3.1 Introduction

Individuals involved in time-to-event studies are monitored for some time after they join the study. For example, in a medical context, researchers will follow up on the patient's health condition over time after treatment. The length of follow-up is an informative component in survival analysis. Korn (1986) noted that, without extrapolation, the results of the analysis should only apply to the time frame in which the individuals in the study were monitored. However, this is also a concept that researchers sometimes ignore. For example, after reviewing 132 papers using survival analysis published in cancer journals, Altman, De Stavola, Love and Stepniewska (1995) noted that about half of them did not include any summary of follow-up time. Independently of Altman et al. (1995), Schemper and Smith (1996) surveyed 70 papers published in medical journals that used survival analysis. They reported that among the 47 out of 70 articles that included a statement regarding the duration of follow-up, 24 of them did not specify the method used to quantify follow-up. In addition, recent reviews by Akl et al. (2012) and Betensky (2015) suggested that researchers are still not paying enough attention to the amount of follow-up.

When analysing survival data with immunes present, it is common to assume that the follow-up in the sample has been sufficient in some sense to allow long-term survivors in a trial to become apparent. However, applying cure models to data without attending to the length of follow-up might lead to misleading or meaningless results. Maller and Zhou (1996) addressed this issue. Their criterion is that there is sufficient follow-up when the censoring time distribution has right extreme τ_G greater than or equal to the right extreme of the distribution of the susceptibles' survival times (with right extreme τ_{F_0}). When a parametric model is fitted

and there is insufficient follow-up, that is, where $\tau_G < \tau_{F_0}$, it will be difficult to distinguish between a high cure proportion with a short tail for the susceptible survival function and a low cure proportion with a long tail (Li, Taylor and Sy, 2001). This problem is often referred to as the "near non-identifiability problem".

In Section 3.2, we demonstrate the effect of the "near non-identifiability problem" on statistical inference using parametric mixture cure models as examples. Researchers have developed various tests and methods to detect insufficient follow-up, and we will review and discuss these in Section 3.3.

3.2 Near non-identifiability problem and flat likelihood

Recall in Section 2.2 we concluded that a mixture cure model with F_0 specified parametrically is nearly always identified under sufficient follow-up. However, it is near non-identifiable under insufficient follow-up, as suggested by Yu, Tiwari, Cronin and Feuer (2004). To better illustrate this problem, we simulate the survival times of $n = 1000$ patients from an exponential distribution with parameter $\lambda = 1$. Each patient is assigned a Bernoulli random variable B_i, whose parameter is $p = 0.75$. The patients with $B_i = 1$ are classified as susceptibles, and the rest with $B_i = 0$ are immunes with actual survival times of infinity. The censoring distribution is assumed to be Uniform$(0, \tau_G)$ with an additional 5% of the censoring times fixed as τ_G. This ensures that the maximum censoring time simulated is exactly τ_G, and it can be interpreted as meaning that 5% of the patients intend to stay until the end of the study unless they experience the event of interest earlier. When there is heavy censoring, this setup nearly always guarantees that a proportion of the censoring times reach τ_G, which is useful for further analysis. In this case, we select τ_G to be

$$q_r + (q_{0.75} - q_{0.25})\,(r_q - 1)\,\mathbf{1}\,(r_q > 1)\,, \tag{3.1}$$

where q_r represents the $(r_q \times 100)-$th percentile of the 1000 simulated actual survival times and $\mathbf{1}(A)$ is a indicator function that equals 1 if the event A is true. The constant r_q controls the amount of censoring. Note that $r_q < 1$ indicates that τ_G will be selected as q_r, which is less than the largest actual survival time simulated. Hence, by definition, it signals insufficient follow-up. In this section, we use $p = 0.75$ and $r_q = 0.8$ for simulation to produce datasets with insufficient follow-up. Though the problem of insufficient follow-up also affects survival

models when there are no immunes, we include a proportion of long-term survivors in this simulation study to emphasise the effect of it on cure models specifically.

The Kaplan-Meier estimate for the simulated survival data is provided in Figure 3.1.

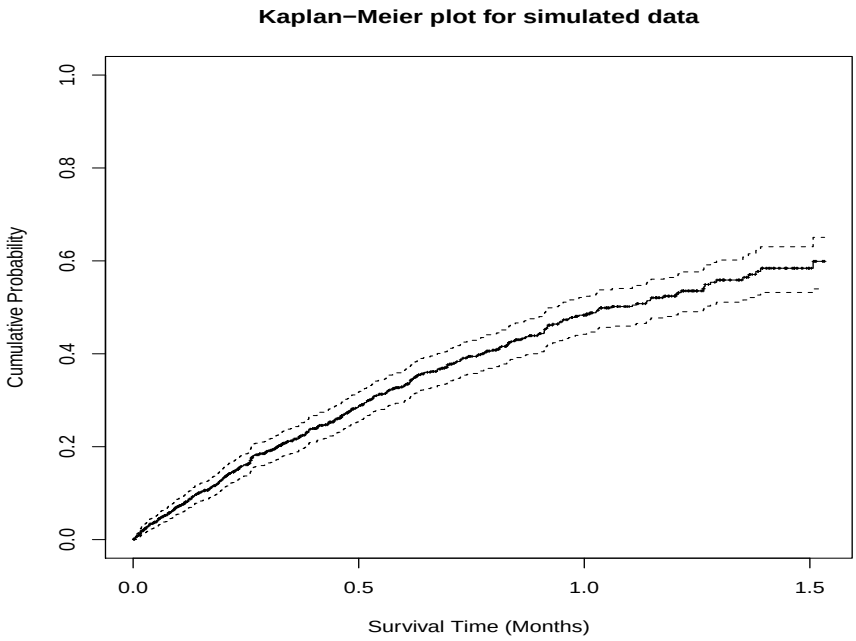


FIGURE 3.1: Plot of Kaplan-Meier estimate for survival data with $n = 1000$ and $r_q = 0.8$. The dotted lines represent the 95% confidence interval.

From the plot, it is unclear whether the curve will level off at the current level (i.e. a high cure proportion with a short tail) or it will continue to increase and then level off at a higher point than the one observed (i.e. a low cure proportion with a long tail). Hence, in this case with $p = 0.75$ and $r_q = 0.8$, the sample has insufficient follow-up and the mixture cure model is non-identifiable. The insufficiency of follow-up can be confirmed by a q_n test (see Section 3.3) with a significance level of $\alpha_s = 0.05$.

We fit an exponential mixture cure model with a distribution function

$$F(t) = p\,[1 - \exp(-\lambda t)]$$

to the simulated data. In Figure 3.2, we present the plot of the profile log-likelihood, $\log L(p, \hat{\lambda}(p))$, against p, where $\log L(p, \hat{\lambda}(p))$ is the maximum log-likelihood value for a given p.

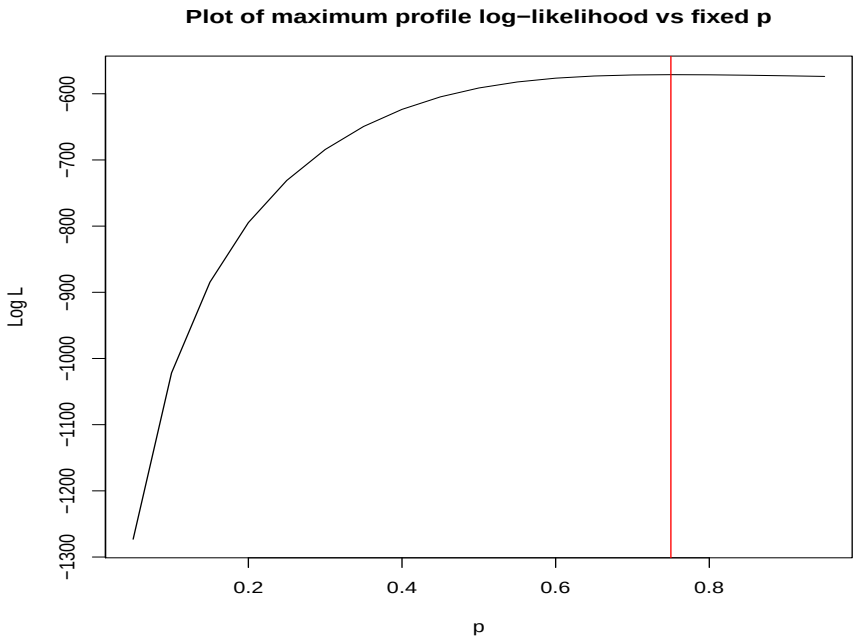


FIGURE 3.2: Plot of maximum profile log-likelihood against fixed susceptible proportion p with $n = 1000$ and $r_q = 0.8$. The vertical line indicates $p = 0.75$.

We observe that the profile log-likelihood is quite flat when $p > 0.6$, which indicates that different sets of parameters $(p, \hat{\lambda}_p)$, with $p > 0.6$, may produce similar log-likelihood values. This results in an unstable maximum likelihood estimator for p with a large variance. In this case, using the "flexsurvcure" package in R (for details, see Section 5.4.1), the maximum likelihood estimator for p, with the correct model specified, is 0.8052 with a very wide 95% confidence interval of $[0.5968, 0.9203]$.

This problem is, of course, more evident with a shorter follow-up period, i.e. smaller r_q. Moreover, it is not remedied with an increased sample size as shown by Figure 3.3, where we changed the sample size n for the simulated data from 1000 to 10000 holding everything else unchanged.

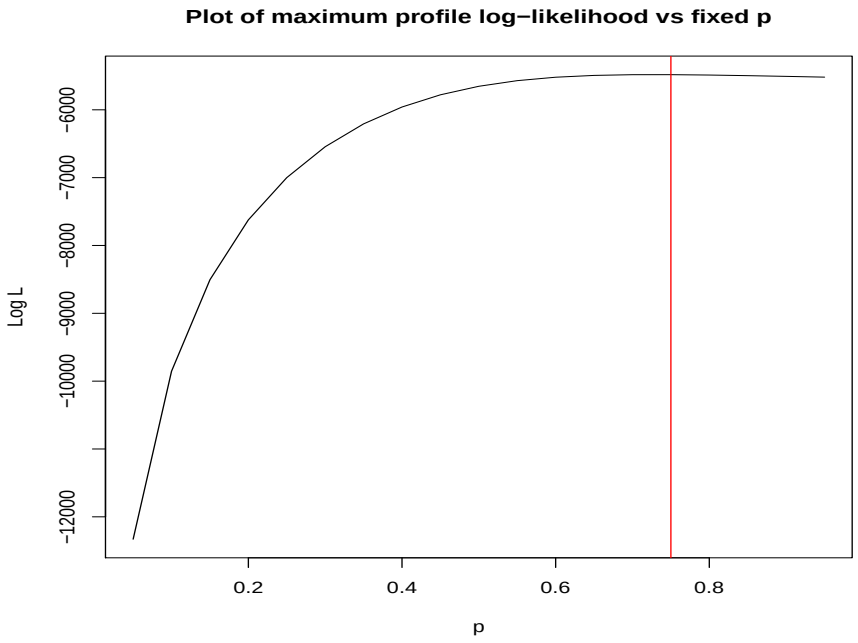


FIGURE 3.3: Plot of maximum profile log-likelihood against fixed susceptible proportion p with $n = 10000$ and $r_q = 0.8$. The vertical line indicates $p = 0.75$.

On the other hand, when there is sufficient follow-up, this problem disappears. For illustration, we simulated another dataset with $n = 1000$, $p = 0.75$, $F_0 \sim \text{Exponential}(1)$ and $r_q = 1$. In this case, we note that the follow-up time should be sufficient with $\tau_G = \tau_{F_0}$. The Kaplan-Meier estimator for the simulated survival data is provided in Figure 3.4. We observe from the plot that the Kaplan-Meier estimator curve levels off at its right tail, which also suggests that the follow-up time is likely to be sufficient.

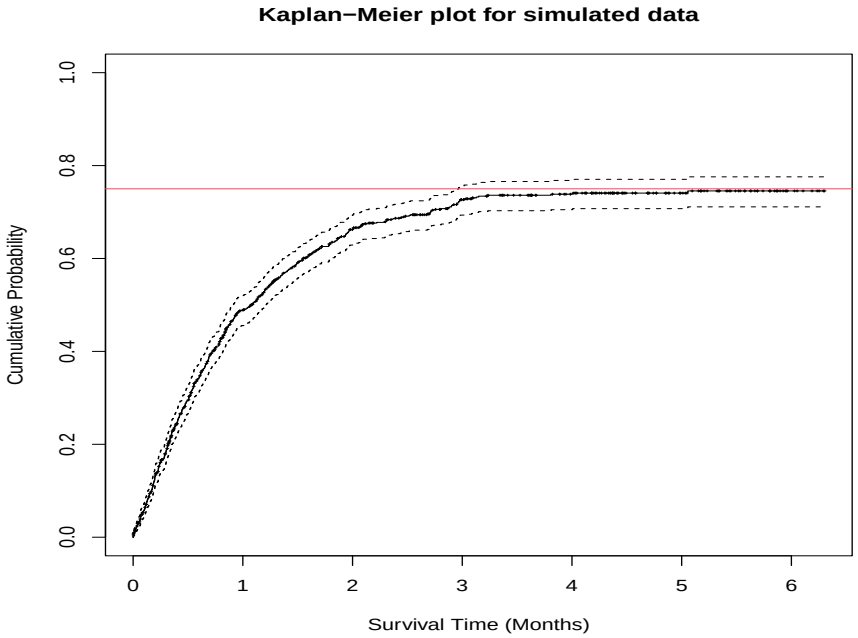


FIGURE 3.4: Plot of Kaplan-Meier estimate for survival data with $n = 1000$ and $r_q = 1$. The dotted lines represent the 95% confidence interval. The red horizontal line represents the true $p = 0.75$.

Figure 3.5 presents the plot of profile log-likelihood, $\log L(p, \hat{\lambda})$, against p, with $n = 1000$, $p = 0.75$ and $r_q = 1$. In this case, the maximum likelihood estimate for p is 0.7504 with a 95% confidence interval $[0.7179, 0.7803]$.

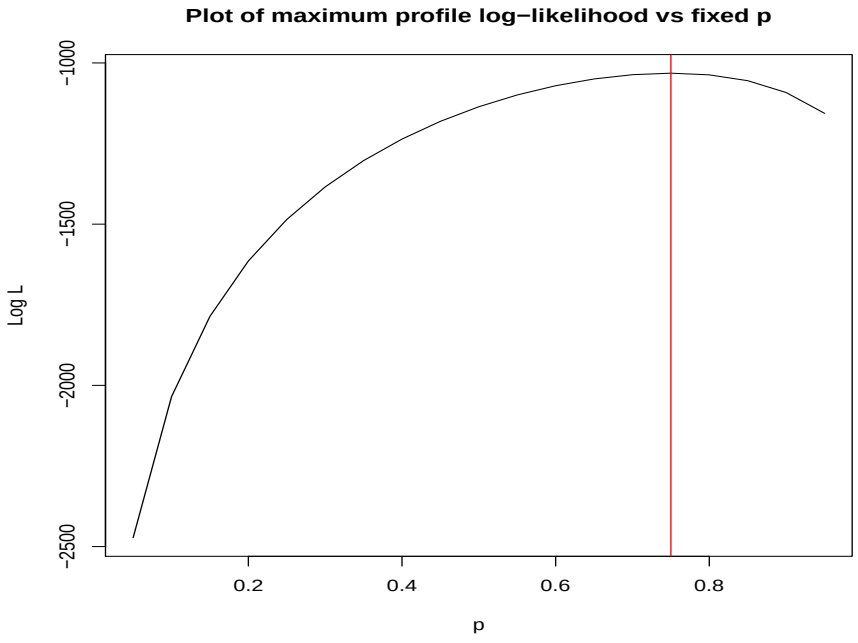


FIGURE 3.5: Plot of maximum profile log-likelihood against fixed susceptible proportion p with $n = 1000$ and $r_q = 1$. The vertical line indicates $p = 0.75$.

Usually, when a model is non-identifiable, our immediate impression may be that the model is overly complicated, see, for example, Huang (2005). However, as demonstrated above with a simple exponential mixture model, the non-identifiability issue present in our case is not due to over-fitting. Instead, it arises under insufficient follow-up due to the lack of information contained in the dataset. This result aligns with and elaborates on what Yu et al. (2004) suggested.

After noting the relationship between insufficient follow-up and a flat likelihood, it is natural to consider the observed information in the sample, which is a measure of the curvature of the log-likelihood function. We start by writing the log-likelihood function, given the observed survival times $\boldsymbol{T}$ and the censoring status $\boldsymbol{C}$, as

$$\begin{aligned} L(p, \Theta; \boldsymbol{T}, \boldsymbol{C}) &= \prod_{i=1}^{n} \left(p f_0(T_i; \Theta)\right)^{C_i} \left(1 - p F_0(T_i; \Theta)\right)^{(1-C_i)}, \\ \ell(p, \Theta; \boldsymbol{T}, \boldsymbol{C}) &= \sum_{i=1}^{n} C_i \log\left(p f_0(T_i; \Theta)\right) + \sum_{i=1}^{n} (1 - C_i) \log\left(1 - p F_0(T_i; \Theta)\right), \end{aligned} \tag{3.2}$$

where Θ is the vector of parameters for the distribution F_0, and f_0 is the corresponding probability density function for F_0. We note that T_i and C_i are realisations of the random variables T and C for the i-th observation respectively.

In this case, only the parameter p is of interest, and the parameters Θ can thus be regarded as nuisance terms. Using (3.2), we compute the score function, the hessian function and the information function for p as

$$\begin{aligned} \text{sc}(p) = \ell'(p) &= \frac{\sum_{i=1}^{n} C_i}{p} + \sum_{i=1}^{n} (1 - C_i) \frac{-F_0(T_i)}{1 - p F_0(T_i)}, \\ \text{hess}(p) = \ell''(p) &= -\frac{\sum_{i=1}^{n} C_i}{p^2} - \sum_{i=1}^{n} (1 - C_i) \frac{F_0(T_i)^2}{\left(1 - p F_0(T_i)\right)^2}. \end{aligned}$$

We now restrict our attention to the total observed information, which is defined as

$$\text{info}(p) = -\ell''(p) = \frac{\sum_{i=1}^{n} C_i}{p^2} + \sum_{i=1}^{n} (1 - C_i) \frac{F_0(T_i)^2}{\left(1 - p F_0(T_i)\right)^2}.$$

Using the relationship $F(t) = p F_0(t)$, the above equation can be written as

$$\text{info}(p) = \frac{\sum_{i=1}^{n} C_i}{p^2} + \sum_{i=1}^{n} (1 - C_i) \frac{F(T_i)^2}{p^2 \left(1 - F(T_i)\right)^2}.$$

To retain the advantages of nonparametric estimation, we estimate $F(t)$ using nonparametric

methods. In this way, we avoid restricting inference to a family of models that can be indexed with a finite-dimensional set of parameters. Let $\hat{F}_n(t)$ be the Kaplan-Meier estimate of the distribution function. The nonparametric estimate of the total observed information is

$$\text{info}_n(p) = \frac{\sum_{i=1}^{n} C_i}{p^2} + \sum_{i=1}^{n} (1 - C_i) \frac{\hat{F}_n(T_i)^2}{p^2 \left(1 - \hat{F}_n(T_i)\right)^2}. \tag{3.3}$$

Using the set-up described at the start of this section, we simulate datasets with $n = 1000$ from an exponential mixture cure model with $\lambda = 1$ and $p = 0.75$. In this case, to visualise the effect of the extent of follow-up on the total observed information, we simulated a dataset for each $r_q \in \{0.10, 0.11, ..., 1.49, 1.50\}$. The nonparametric estimate of the total observed information, which is computed using the true $p = 0.75$, is then plotted against the level of follow-up r_q.

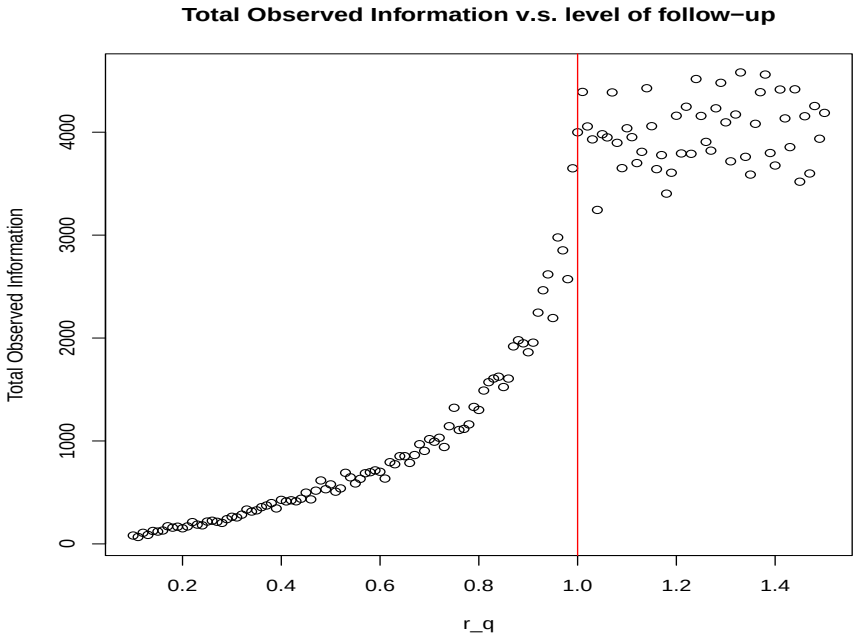


FIGURE 3.6: Plot of nonparametric estimate of the total observed information $\text{info}_n(0.75)$ against the level of follow-up r_q with $n = 1000$, $\lambda = 1$ and $p = 0.75$.

From Figure 3.6, we note that when the follow-up time is sufficient, that is when $r_q \geq 1$, the information contained in the dataset for the parameter p is relatively stable. However, it decreases significantly with decreasing r_q under insufficient follow-up. Overall, we conclude that the total observed information increases with the level of follow-up time, and will stay relatively constant once the follow-up period becomes sufficient. This supports our argument that the problem of insufficient follow-up is due to a lack of information, which is caused by the increased proportion of censored observations in the data.

3.3 Review of tests for sufficient follow-up

The problem of insufficient follow-up has a huge impact on statistical inference, it is natural for us to ask: given a set of survival data, how do we decide whether it has sufficient follow-up? Only a few papers have addressed the issue and proposed ideas for solving this problem. However, all of them have their limitations.

3.3.1 Maller and Zhou's q_n test

Maller and Zhou (1992) defined insufficient follow-up as occurring when $\tau_{F_0} > \tau_G$, and based on this definition, Maller and Zhou (1996) proposed a hypothesis test

$$H_0 : \tau_{F_0} > \tau_G \quad \text{(Insufficient follow-up)},$$
$$H_1 : \tau_{F_0} \leq \tau_G \quad \text{(Sufficient follow-up)},$$

with the proposed test statistic

$$q_n = \frac{N_n}{n},$$

where n is the sample size and N_n is the number of uncensored observations in $\left(2\tau_{(k)} - t_{(n)}, \tau_{(k)}\right]$, with $\tau_{(k)}$ being the largest uncensored survival time. Maller and Zhou (1996) showed that under some reasonable conditions, q_n converges in probability to 0 asymptotically when follow-up is insufficient. In other words, there is evidence to conclude that the follow-up period is insufficient if q_n is less than a critical value. These values are tabulated in Maller and Zhou (1996) for a range of sample sizes and censoring distributions, while the survival times are assumed to be exponentially distributed. An application of this test can be found in Section 5.5.1. More details on this hypothesis test can also be found in Chapter 4.3 of Maller and Zhou (1996).

Klebanov and Yakovlev (2007) noted that the above q_n test's critical values are decided using simulations based on parametric assumptions on F and G. They commented that the above hypothesis test is not conclusive as the parametric assumption for the survival times deprived the proposed test of all non-parametric advantages. Maller, Resnick and Shemehsaver (2021a) recently addressed this issue. They investigated the test statistic q_n and derived its distribution under some conditions. They showed that in a finite sample, assuming the independent and identically distributed censoring model and using the notation of Section 1.3, for $n > 2$ and

$k_q = 0, 1, 2, ..., n-2,$

$$P\left(nq_n = k_q \mid 0 < \tau_{(k)} < t_{(n)}\right) = \frac{A_n(k_q) + B_n(k_q)}{D_n}, \tag{3.4}$$

where $t_{(n)}$ is the largest survival time observed,

$$A_n(k_q) = \int_{t=0}^{\tau_H/2} \int_{x=2t}^{\tau_H} P\left(\text{Bin}\left(n-2, \rho^A(t,x)\right) = k\right) P_n(\,\mathrm{d}t,\,\mathrm{d}x),$$

with $\tau_H = \min(\tau_G, \tau_{F_0})$, and

$$B_n(k_q) = \left[\int_{t=0}^{\tau_H/2} \int_{x=t}^{2t} + \int_{t=\tau_H/2}^{\tau_H} \int_{x=t}^{\tau_H}\right] P\left(\text{Bin}\left(n-2, \rho^B(t,x)\right) = k\right) P_n,$$

for certain quantities ρ^A, ρ^B, $P_n(\,\mathrm{d}t,\,\mathrm{d}x)$ and D_n such that

$$\begin{aligned}
\rho^A(t,x) &= \left(1 - \frac{\int_{y=t}^{x} \bar{F}(y)\mathrm{d}G(y)}{\int_{y=t}^{x} \bar{F}(y)\mathrm{d}G(y) + H(t)}\right) \frac{\int_0^t \bar{G}(y)\mathrm{d}F(y)}{H(t)}, \\
\rho^B(t,x) &= \left(1 - \frac{\int_{y=t}^{x} \bar{F}(y)\mathrm{d}G(y)}{\int_{y=t}^{x} \bar{F}(y)\mathrm{d}G(y) + H(t)}\right) \frac{\int_{2t-x}^{t} \bar{G}(y)\mathrm{d}F(y)}{H(t)}, \\
P_n(\,\mathrm{d}t,\,\mathrm{d}x) &= n(n-1)\left(\int_{y=t}^{x} \bar{F}(y)\mathrm{d}G(y) + H(t)\right)^{n-2} \bar{G}(t)\mathrm{d}F(t)\bar{F}(x)\mathrm{d}G(x), \\
D_n &= 1 - \left(\int_{t=0}^{\tau_H} \bar{F}(z)\mathrm{d}G(z)\right)^n - n\int_{t=0}^{\tau_H} H^{n-1}(t)\bar{G}(t)\mathrm{d}F(t),
\end{aligned}$$

with $H(t) := P(T \leq t)$ being the distribution of the observed survival times. Here $\overline{F}(t) = 1 - F(t)$ denotes the survival function of F; similarly, for F_0, G and H.

Assume the independent and identically distributed censoring model (see Section 1.3), $0 < p \leq 1$, $\tau_G < \tau_{F_0} \leq \infty$ and

$$\bar{G}\left(\tau_G - z\right) = a_G(1 + o(1))z \text{ as } z \downarrow 0,$$

for a constant $a_G > 0$. Suppose also F_0 has a density in a neighbourhood of τ_G, which is positive and continuous at τ_G. Then when $n \to \infty$, the authors showed that nq_n is asymptotically geometric with parameter 0.25, that is

$$\lim_{n\to\infty} P(nq_n = k_q) = \frac{1}{4}\left(\frac{3}{4}\right)^{k_q}, \quad k_q = 0, 1, 2, \cdots$$

Therefore, the critical value for the q_n test when n is large is the $100(1-\alpha_s)\%$ percentile of the above geometric distribution. For example, when $\alpha_s = 0.05$, the critical value is 10.

Applying the q_n test in practice requires the probability mass function (3.4) for the test statistics nq_n in a finite sample. To compute (3.4), we need specific assumptions for $F = pF_0$ and G that validate the null hypothesis (i.e. insufficient follow-up). As an example, when we apply the q_n test in practice, our choices can be $p = 0.75$ and F_0 being the distribution function of an exponential distribution with parameter $\lambda = 1$. We assume the censoring distribution G is a uniform distribution between 0 and $\tau_G = 1.5$. The sample size is chosen to be $n = 50$. We note that the 99% percentile of F_0 is approximately 4.61. We choose $\tau_G = 1.5$ that is smaller than 4.61. Under this particular set-up, for all $n = 50$ observations, the probability of their maximum actual survival time (τ_{F_0}) being smaller than or equal to the follow-up time (τ_G) is

$$\begin{aligned} P(H_0 \text{ is false}) = P(\tau_{F_0} \leq \tau_G) &= \prod_{i=1}^{50} P(T_i^* \leq \tau_G) \\ &= (1 - \exp(-\tau_G))^{50} \\ &= 3.2916 \times 10^{-6} \approx 0. \end{aligned}$$

Therefore, our choices of $n = 50$, $p = 0.75$, $F_0 \sim \text{Exponential}(1)$ and $G \sim \text{Uniform}(0, 1.5)$ meet the requirement for insufficient follow-up.

After computing the pmf for nq_n using (3.4), we find the $100(1-\alpha_s)\%$ percentile of nq_n, the critical value for the hypothesis test. In our case, when the significance level $\alpha_s = 0.05$, the critical value is 13 because $P(nq_n \leq 13) = 0.95$. In other words, we will reject the null hypothesis and conclude the data has sufficient follow-up if the observed nq_n is larger than 13 under the above assumptions.

Klebanov and Yakovlev (2007) commented that it is possible for a censored susceptible individual to experience the event of interest if the follow-up period is extended beyond the current level. In this case, the value of q_n calculated on the sample with extended follow-up will decrease from the former value. They view this non-monotone behaviour of q_n as a disadvantage. On the other hand, Maller, Resnick and Shemehsaver (2021) argued that at each stage, q_n is correctly indicating the extent of follow-up, and there is no reason for q_n to be strictly monotone in a hypothetical situation of increasing follow-up.

Maller and Zhou (1996, p.g. 85) showed that information on the difference $\tau_G - \tau_{F_0}$ is provided by the distance between $t_{(n)}$, the largest observed survival or censoring time, and τ_k, the largest observed survival time, at least when n is large. They proposed the q_n test and the non-parametric test statistic N_n based on that idea, where a large N_n usually indicates a large $t_{(n)} - \tau_k$ and hence $\tau_G - \tau_{F_0}$ when n is large. At the same time, we note that the power of the q_n test also depends on the difference $\tau_G - \tau_{F_0}$. That is, when $\tau_G - \tau_{F_0}$ is non-negative but small, the q_n test might not be as powerful. This issue is addressed in Maller, Resnick and Shemehsaver (2021), where the authors showed that the power of the q_n test increases as τ_G increases above τ_{F_0} when $\tau_{F_0} < \tau_G < 2\tau_{F_0}$.

Maller, Resnick and Shemehsaver (2021) also noted that the q_n test is sensitive to the occurrence of one or few uncensored observations at the right end of the Kaplan-Meier estimator and this robustness issue has to be addressed in statistical analysis. They suggested using a test for outliers in the independent and identically distributed censoring model, as proposed in Maller and Zhou (1994).

3.3.2 Shen (2000) test

Following a similar idea to Maller and Zhou (1996), Shen (2000) modified the q_n statistic to construct the so-called $\tilde{\alpha}_n$ test. As mentioned above, it is obvious that when $p < 1$ and $\tau_G - \tau_{F_0}$ is sufficiently large, the Kaplan-Meier plot should level-off at $\hat{p}_n$ for $t \in (\tau_{(k)}, t_{(n)}]$. Shen (2000) refers to the levelling as the "final interval of constancy". Similar to the q_n test, Shen's test is constructed based on the idea that there will be evidence for insufficient follow-up if this final interval of constancy is not large enough. Equivalently, Shen (2000) suggested that the null hypothesis $H_0 : \tau_{F_0} > \tau_G$ should be rejected if the observed $r_{(n)} = t_{(n)}/\tau_{(k)}$ is too large. Denote the test statistic as $T_{(n)}/T^*_{(k)}$, whose realisation is $r_{(n)} = t_{(n)}/\tau_{(k)}$, the p-value of Shen's proposed hypothesis test is hence

$$P(T_{(n)}/T^*_{(k)} > r_{(n)}),$$

which is bounded above by

$$P(T^*_{(k)} < \tau_G/r_{(n)}),$$

because $T_{(n)} \leq \tau_G$. Shen (2000) then suggested estimating τ_G using $\hat{\tau}_G = \hat{w}\tau_{(k)} + (1-\hat{w})t_{(n)}$, where $\hat{w} = (t_{(n)} - \tau_{(k)})/t_{(n)}$. Using $\hat{\tau}_G$, the upper bound for the p-value can be estimated using

$$\tilde{\alpha}_n = (1-\tilde{q}_n)^n,$$

with

$$\tilde{q}_n = \frac{\tilde{N}_n}{n}, \tag{3.5}$$

and $\tilde{N}_n$ equals to the number of uncensored observations in $[\hat{\tau}_G\tau_{(k)}/t_{(n)}, \tau_{(k)}]$.

Based on simulations, Shen concluded that both the q_n test and the $\tilde{\alpha}_n$ test have a substantially smaller distortion of nominal significance level than the $\hat{\alpha}_n$ test [1]. We note that Shen's test for sufficient follow-up is based on the length of the interval between $t_{(n)}$ and τ_k. Therefore, just like the q_n test, the power of Shen's test is also dependent on τ_G. In addition, Shen (2000) did not provide a formula for the distribution of the test statistic $\tilde{\alpha}_n$. Further research is needed to investigate the properties of Shen's statistic and how they compare with those of q_n.

3.3.3 Klebanov and Yakovlev (2007) test

Klebanov and Yakovlev (2007) proposed a test for sufficient follow-up using a different idea from the q_n test and the $\tilde{\alpha}_n$ test. They commented that when performing a test for sufficient follow-up, what the researchers "would like to know is whether the follow-up has been long enough to provide strong evidence for the presence of immunes in the study population". Based on this idea, they proposed a method to derive the lower bound of the cure proportion $1-p$ and used it to test

$$H_0 : p = 1$$
$$H_1 : p < 1.$$

By doing so, Klebanov and Yakovlev (2007) noted that they changed the concept of sufficient follow-up. Namely, they disregard the condition $\tau_G - \tau_{F_0} \geq 0$ while focusing solely on their hypotheses [2].

[1] Before proposing the q_n test, Maller and Zhou (1992,1994) initially suggested using $\hat{\alpha}_n = (1-q_n)^n$, a monotone transformation of q_n, as the estimates for the p-values. They later discovered the so-called $\hat{\alpha}_n$ test is incorrect. They acknowledged this explicitly in Maller and Zhou (1996, p.g. 84).

[2] López Segovia, L. (2014) states in his/her thesis that Klebanov and Yakovlev's set of hypotheses is equivalent to the one proposed by Maller and Zhou (1996), which is not correct.

Klebanov and Yakovlev (2007) aimed to construct a lower confidence limit for the difference $S(\tau_G) - S_0(\tau_G)$ under the assumption that the survival function for the susceptibles (the latency model) has non-decreasing φ-hazard rate average, where the φ-hazard rate for a model with survival function S is defined as

$$r(t) = \frac{\mathrm{d}}{\mathrm{d}t}\varphi^{-1}(S(t)),$$

with a non-negative strictly monotonically decreasing function $\varphi(u)$, whose first derivative φ' is continuous and $\varphi(0) = -\varphi'(0) = 1$.

Under the above assumption, Klebanov and Yakovlev (2007) computed a lower bound for $S(\tau_G) - S_0(\tau_G)$ as

$$\Delta(\tau_G) = S(\tau_G) - \varphi\left(\frac{\tau_G}{t_0}\varphi^{-1}\left(S\left(t_0\right)\right)\right), \quad t_0 \in (0, \tau_G). \tag{3.6}$$

Then they made another "mild" assumption that the function

$$\psi(u) = \varphi\left(\frac{\tau_G}{t_0}\varphi^{-1}(u)\right), \quad 0 \leqslant t_0 \leqslant \tau_G$$

is convex in u. The authors noted that this assumption is met in the special case of the traditional hazard function, where $\varphi(u) = \exp(-u)$. Under this assumption and using the theory for the Kolmogorov goodness-of-fit statistic, Klebanov and Yakovlev (2007) showed that

$$P\left\{\Delta(\tau_G) \geqslant \frac{n-k}{n} - \varphi\left(\frac{\tau_G}{t_0}\varphi^{-1}\left(S_n\left(t_0\right)\right)\right) - \left[1 + \frac{\tau_G}{t_0}\right]\frac{D_{\alpha_s}(n)}{n^{1/2}}\right\} \geqslant 1 - \alpha_s, \tag{3.7}$$

where k is the total number of uncensored observations, $S_n(t)$ is the empirical counterpart of the survival function $S(t)$, α_s is the significance level and $D_{\alpha_s}(n)$ is the $(1-\alpha_s)$-th percentile of the Kolmogorov distribution for a sample of size n. The statistic $\Delta(\tau_G)$ is given by (3.6). Overall, the Klebanov and Yakovlev (2007) test rejects the hypothesis $S(\tau_G) = S_0(\tau_G)$ if the lower confidence limit in (3.7) is greater than 0 at a significance level of less than α_s.

Maller, Resnick and Shemehsaver (2021) noted that Klebanov and Yakovlev (2007) test provides a potentially useful perspective on the problem of insufficient follow-up. However, they commented that Klebanov and Yakovlev's test statistic is "technically complex" and "non-intuitive", whereas a test based on the distance $t_{(n)} - \tau_k$ is more interpretable. Moreover, another significant constraint of Klebanov and Yakovlev (2007) test is that they made the "quite

plausible" basic assumption, that the survival function for the susceptibles (the latency model) has a non-decreasing φ-hazard rate average. This assumption restricts the application of their method. As an example, when the latency model has a Weibull distribution with parameters $\{\lambda, \alpha\}$, and the non-negative strictly monotonically decreasing function is chosen to be $\varphi(u) = \exp(-u)$ with $\varphi^{-1}(x) = -\log(x)$, the corresponding φ-hazard rate is

$$\begin{aligned} r(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\phi^{-1}\left(\exp\left(-\left(\lambda t\right)^{\alpha}\right)\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\lambda^{\alpha}t^{\alpha} \\ &= \alpha\lambda^{\alpha}t^{\alpha-1}, \end{aligned}$$

which is not a non-decreasing function when $\alpha < 1$.

3.3.4 Other methods to detect insufficient follow-up

As an alternative to using a hypothesis test, some authors have tried to estimate the length of the follow-up period required to be sufficient, but their methods are rather informal mathematically. For example, Yu, Tiwari, Cronin and Feuer (2004) suggest that follow-up times need to be at least two-thirds of the median survival time in the uncured patients. Tai, Yu, Cserni, Vlastos, Royce, Kunkler and Vinh-Hung (2005) suggested fitting a log-normal model with parameters μ and σ^2 to the survival data where appropriate to obtain the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}$. They suggested that the minimum follow-up period required is $\hat{\mu}\hat{\sigma}^2$. We note that in their work, this value was interpreted as a "threshold time" for a statistical cure, i.e. observations with survival times larger than the threshold time are considered as cured, so in other words, $\hat{\mu}\hat{\sigma}^2$ is an estimate of τ_{F_0}.

Chapter 4

Parametric Cure Models

4.1 Introduction

As discussed in Chapter 2, in practice, mixture cure models are widely used because they are both conceptually and computationally simple. They also allow us to interpret the effects of covariates on the latency and incidence models separately. Moreover, the non-mixture cure model assumes the overall population follows a proportional hazards structure, which might be violated in practice. Therefore, for the rest of this chapter and this thesis, we will focus on mixture cure models and, specifically, parametric mixture cure models.

The first challenge faced by many researchers when constructing a parametric mixture cure model is to specify a latency model that does not over-fit or under-fit the data. A starting point would be to find a single parametric model that embeds most of the suitable parametric lifetime distributions, so we can carry out nested likelihood ratio tests to select the most appropriate model. In this section, we will work with the flexible four-parameter generalised F distribution, which was introduced by Prentice (1975). We note that, to the best of our knowledge, the literature lacks a clear outline of the relationships between the generalised F model and its sub-models, and we aim to fill this gap in this chapter.

In Sections 4.2, 4.3 and 4.4, we construct the generalised F distribution and derive its connections with lifetime distributions in the generalised gamma family, the generalised logistic family, and a few parametric distributions used in extreme value theory.

4.2 The generalised F distribution and its sub-models

4.2.1 Construction of the generalised F distribution

With the notation defined in Section 1.3, consider a location and scale model with the errors following the logarithm of a F distribution with degrees of freedom $2s_1 > 0$ and $2s_2 > 0$, that is

$$\log(T^*) = \mu + \sigma\varepsilon, \quad T^* \geq 0, \varepsilon \in \mathbb{R},$$

with location parameter μ, scale parameter $\sigma > 0$, and $\exp(\varepsilon) \sim F(2s_1, 2s_2)$. The model can also be written as

$$\exp(\varepsilon) = \exp\left(\frac{\log(T^*) - \mu}{\sigma}\right) \sim F(2s_1, 2s_2), \tag{4.1}$$

where the probability density function for $\exp(\varepsilon)$ is

$$f_{\exp(\varepsilon)}(x; s_1, s_2) = \frac{\Gamma(s_1 + s_2)}{\Gamma(s_1)\Gamma(s_2)} \left(\frac{s_1}{s_2}\right)^{s_1} x^{s_1 - 1} \left(1 + \frac{s_1}{s_2} x\right)^{-s_1 - s_2}.$$

Note the support for the F distribution function is $\exp(\varepsilon) \in [0, +\infty)$. We obtain the probability density function for ε as

$$\begin{aligned} f_\varepsilon(\varepsilon; s_1, s_2) &= f_{\exp(\varepsilon)}(x = \exp(\varepsilon); s_1, s_2) \left|\frac{\mathrm{d}\exp(\varepsilon)}{\mathrm{d}\varepsilon}\right| \\ &= \frac{\Gamma(s_1 + s_2)}{\Gamma(s_1)\Gamma(s_2)} \left(\frac{s_1}{s_2}\right)^{s_1} \exp(\varepsilon)^{s_1 - 1} \left(1 + \frac{s_1}{s_2}\exp(\varepsilon)\right)^{-s_1 - s_2} \exp(\varepsilon) \\ &= \frac{\Gamma(s_1 + s_2)}{\Gamma(s_1)\Gamma(s_2)} \left(\frac{s_1}{s_2}\right)^{s_1} \exp(\varepsilon)^{s_1} \left(1 + \frac{s_1}{s_2}\exp(\varepsilon)\right)^{-s_1 - s_2}. \end{aligned} \tag{4.2}$$

We will interpret ε as a random variable following the two-parameter generalised F distribution with positive parameters s_1 and s_2, where $\varepsilon \in \mathbb{R}$. This distribution is also known in the literature as the log F distribution. Further extending it with location ($\mu \in \mathbb{R}$) and scale ($\sigma \in (0, +\infty)$) parameters results in a four-parameter generalised distribution, which will be referred to as the generalised F distribution for the rest of this thesis. Assuming T^*, the actual survival time for the susceptibles, is a non-negative random variable following such a distribution, its probability density function can be derived using (4.1). In this case, we treat ε as the function of transformation, so

$$\varepsilon(t) = \frac{\log(t) - \mu}{\sigma},$$

and

$$\left|\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}\right| = \frac{1}{t\sigma},$$

so that

$$\begin{aligned} f_{GF}(t;\mu,\sigma,s_1,s_2) &= f_\varepsilon\left(\varepsilon = \frac{\log(t)-\mu}{\sigma}\right)\left|\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}\right| \\ &= \frac{\Gamma(s_1+s_2)}{t\sigma\Gamma(s_1)\Gamma(s_2)}\left(\frac{s_1}{s_2}\right)^{s_1}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)^{s_1}\left(1+\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)\right)^{-s_1-s_2}. \end{aligned} \tag{4.3}$$

4.2.2 The generalised gamma family

In Section 1.4.2, we introduced the generalised gamma distribution, which contains as sub-models a number of distributions that are naturally suitable for modelling lifetime data. Recall that (1.1) defines the probability density function of the generalised gamma distribution with non-negative parameters α, λ and γ as

$$f_{gg}(t) = F_0'(t) = \frac{\alpha}{\Gamma(\gamma)}t^{\alpha\gamma-1}\lambda^{\alpha\gamma}e^{-(\lambda t)^\alpha}, \quad t \geq 0$$

and its cumulative distribution function is

$$\begin{aligned} F_{gg}(t) &= \int_0^{(\lambda t)^\alpha}\frac{1}{\Gamma(\gamma)}z^{\gamma-1}\exp(-z)\mathrm{d}z \\ &:= I\left(\gamma,(\lambda t)^\alpha\right), \end{aligned}$$

where

$$I(\gamma,x) = \frac{1}{\Gamma(\gamma)}\int_0^x u^{\gamma-1}\exp(-u)\mathrm{d}u.$$

Examples of sub-models contained in the generalised gamma distribution were given in Table 1.1.

In addition, recall the extended generalised gamma distribution obtained by replacing γ with $Q = \gamma^{-1/2}$, whose probability density function is defined in (1.5) as

$$f_{T^*}(t) = \frac{|Q|}{t\sigma_e\Gamma(Q^{-2})}(Q^{-2})^{Q^{-2}}\exp\left[Q^{-2}\left(Q\frac{\log t-\mu_e}{\sigma_e} - \exp\left(Q\frac{\log t-\mu_e}{\sigma_e}\right)\right)\right],$$

where $\mu_e = (\log\gamma)/\alpha - \log\lambda$ and $\sigma_e = (\alpha\sqrt{\gamma})^{-1}$.

Recall from Section 1.4.2 that if the non-negative survival time for the susceptibles, T^*, follows the extended generalised gamma distribution, then for the transformation of T^*

$$Z = \frac{\log T^* - \mu_e}{\sigma_e},$$

with its density function f_Z defined by (1.3), $Q < 0$ indicates that $-Z$ has the probability density function $f_Z(z;\gamma)$, $Q = 0$ indicates that Z has a standard normal distribution, and $Q > 0$ indicates that Z has the probability density function $f_Z(z;\gamma)$.

Many authors, including Peng, Dear and Denham (1998), mentioned that the generalised F distribution with probability density function (4.3) reduces to the extended generalised gamma distribution when either s_1 or s_2 approaches infinity. However, we note such a statement is not precise. In this section, we show that when $s_1 \to \infty$, the generalised F distribution reduces to the extended generalised gamma distribution with $Q < 0$; when $s_2 \to \infty$, it reduces to the extended generalised gamma distribution with $Q > 0$.

For demonstration, we first show that when $s_2 \to \infty$,

$$\begin{aligned}
\lim_{s_2\to\infty} f_{GF}(t) &= \lim_{s_2\to\infty} \frac{\Gamma(s_1+s_2)}{t\sigma\Gamma(s_1)\Gamma(s_2)} \left(\frac{s_1}{s_2}\right)^{s_1} \exp\left(\frac{\log(t)-\mu}{\sigma}\right)^{s_1} \times \\
&\qquad \left(1+\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)\right)^{-s_1-s_2} \\
&= \frac{s_1^{s_1}}{t\sigma\Gamma(s_1)} \exp\left(\frac{\log(t)-\mu}{\sigma}\right)^{s_1} \lim_{s_2\to\infty} \frac{\Gamma(s_1+s_2)}{\Gamma(s_2)s_2^{s_1}} \left(1+\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)\right)^{-s_1-s_2} \\
&= \frac{s_1^{s_1}}{t\sigma\Gamma(s_1)} \exp\left(\log(t)\right)^{s_1/\sigma} \exp\left(-\mu\right)^{s_1/\sigma} \lim_{s_2\to\infty} \frac{\Gamma(s_1+s_2)}{\Gamma(s_2)s_2^{s_1}} \times \\
&\qquad \left(1+\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)\right)^{-s_1-s_2} \\
&= \frac{\sigma^{-1}}{\Gamma(s_1)} t^{s_1/\sigma-1} \left(\exp\left(-\mu\right) s_1^{\sigma}\right)^{s_1/\sigma} \lim_{s_2\to\infty} \frac{\Gamma(s_1+s_2)}{\Gamma(s_2)s_2^{s_1}} \times \\
&\qquad \lim_{s_2\to\infty} \left(1+\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)\right)^{-s_1-s_2}. \qquad (4.4)
\end{aligned}$$

Using Stirling's approximation for the gamma function

$$\Gamma(x) = (2\pi)^{1/2} x^{x-\frac{1}{2}} \exp(-x) \left(1 + O\left(\frac{1}{x}\right)\right),$$

with $O\left(\frac{1}{x}\right)$ as the error term, we can approximate the ratio of gamma functions. Note that as $s_2 \to \infty$, the error $O\left(\frac{1}{s_2}\right)$ can be ignored. We get

$$\begin{aligned}
\lim_{s_2\to\infty} \frac{\Gamma(s_1+s_2)}{\Gamma(s_2)s_2^{s_1}} &= \lim_{s_2\to\infty} \frac{(s_1+s_2)^{(s_1+s_2)}(s_1+s_2)^{-1/2}\exp(-(s_1+s_2))}{s_2^{s_2}s_2^{-1/2}\exp(-s_2)s_2^{s_1}} \\
&= \lim_{s_2\to\infty} \left(\frac{s_1+s_2}{s_2}\right)^{s_2}\left(\frac{s_1+s_2}{s_2}\right)^{s_1}\left(\frac{s_1+s_2}{s_2}\right)^{-\frac{1}{2}} \frac{\exp(-s_1)\exp(-s_2)}{\exp(-s_2)} \\
&= \exp(-s_1)\lim_{s_2\to\infty}\left(1+\frac{s_1}{s_2}\right)^{s_2}\lim_{s_2\to\infty}\left(1+\frac{s_1}{s_2}\right)^{s_1-\frac{1}{2}} \\
&= \exp(-s_1)\lim_{s_2\to\infty}\left(1+\frac{s_1}{s_2}\right)^{s_2}. \qquad (4.5)
\end{aligned}$$

Using

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = \exp(x),$$

it is then easy to see that (4.5) becomes

$$\lim_{s_2\to\infty} \frac{\Gamma(s_1+s_2)}{\Gamma(s_2)s_2^{s_1}} = \exp(-s_1)\exp(s_1) = 1. \qquad (4.6)$$

With (4.2.2), we also obtain that

$$\begin{aligned}
\lim_{s_2\to\infty}\left(1+\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)\right)^{-s_1-s_2} &= \lim_{s_2\to\infty}\left(1+\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)\right)^{-s_1} \times \\
&\quad \lim_{s_2\to\infty}\left(1+\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)\right)^{-s_2} \\
&= \lim_{s_2\to\infty}\left(1+\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)\right)^{-s_2} \\
&= \exp\left(-s_1\exp\left(\frac{\log(t)-\mu}{\sigma}\right)\right) \\
&= \exp\left(-s_1 t^{1/\sigma}\exp\left(-\mu\right)^{1/\sigma}\right) \\
&= \exp\left(-\left(s_1^{\sigma}\exp\left(-\mu\right)t\right)^{1/\sigma}\right). \qquad (4.7)
\end{aligned}$$

Substituting (4.6) and (4.7) into (4.4), we obtain

$$\lim_{s_2\to\infty} f_{GF}(t) = \frac{\sigma^{-1}}{\Gamma(s_1)} t^{s_1/\sigma-1}\left(\exp\left(-\mu\right)s_1^{\sigma}\right)^{s_1/\sigma}\exp\left(-\left(s_1^{\sigma}\exp\left(-\mu\right)t\right)^{1/\sigma}\right). \qquad (4.8)$$

The limit in (4.8) is the probability density function of a generalised gamma distribution (1.1) with parameters $\alpha = \sigma^{-1}$, $\lambda = s_1^{\sigma}\exp(-\mu)$ and $\gamma = s_1$.

As discussed previously, the generalised gamma distribution with parameters $(\alpha, \lambda, \gamma)$ can be transformed to the extended generalised gamma distribution with parameters (μ_e, σ_e, Q) through the reparameterisation

$$\mu_e = \log(\gamma)/\alpha - \log(\lambda), \quad \sigma_e = (\alpha\sqrt{\gamma})^{-1}, \quad Q^{-2} = \gamma. \tag{4.9}$$

In this case, relating the parameters of the extended generalised gamma distribution to those of the generalised F distribution results in

$$\begin{aligned} \mu_e &= \log(\gamma)/\alpha - \log(\lambda) & \sigma_e &= (\alpha\sqrt{\gamma})^{-1} & Q &= \sqrt{\gamma}^{-1/2} \\ &= \sigma\log(s_1) - \sigma\log(s_1) + \mu & &= \frac{\sigma}{\sqrt{s_1}}; & &= \sqrt{s_1}^{-1}. \\ &= \mu; \end{aligned}$$

We note that s_1 should be strictly positive and consequently, $Q > 0$.

Alternatively, if we consider the case where $s_1 \to \infty$ instead, the generalised F model with probability density function (4.3) becomes

$$\begin{aligned} \lim_{s_1\to\infty} f_{GF}(t) &= \lim_{s_1\to\infty} \frac{\Gamma(s_1+s_2)}{t\sigma\Gamma(s_1)\Gamma(s_2)} \left(\frac{s_1}{s_2}\right)^{s_1} \exp\left(\frac{\log(t)-\mu}{\sigma}\right)^{s_1} \left(1+\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)\right)^{-s_1-s_2} \\ &= \frac{1}{t\sigma\Gamma(s_2)} \lim_{s_1\to\infty} \frac{\Gamma(s_1+s_2)}{\Gamma(s_1)} \left(\frac{\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)}{1+\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)}\right)^{s_1} \times \\ &\qquad \left(1+\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)\right)^{-s_2}. \end{aligned} \tag{4.10}$$

Note that similar to (4.5), we have

$$\begin{aligned} \lim_{s_1\to\infty} \frac{\Gamma(s_1+s_2)}{\Gamma(s_1)} &= \lim_{s_1\to\infty} \frac{(s_1+s_2)^{(s_1+s_2)}(s_1+s_2)^{-1/2}\exp(-(s_1+s_2))}{s_1^{s_1}s_1^{-1/2}\exp(-s_1)} \\ &= \lim_{s_1\to\infty} \left(\frac{s_1+s_2}{s_1}\right)^{s_1} (s_1+s_2)^{s_2} \left(\frac{s_1+s_2}{s_1}\right)^{-\frac{1}{2}} \frac{\exp(-s_1)\exp(-s_2)}{\exp(-s_1)} \\ &= \exp(-s_2) \lim_{s_1\to\infty} \left(1+\frac{s_2}{s_1}\right)^{s_1} \lim_{s_1\to\infty} \left(1+\frac{s_2}{s_1}\right)^{-\frac{1}{2}} \lim_{s_1\to\infty} (s_1+s_2)^{s_2} \\ &= \exp(-s_2)\exp(s_2) \lim_{s_1\to\infty} (s_1+s_2)^{s_2} \\ &= \lim_{s_1\to\infty} (s_1+s_2)^{s_2}. \end{aligned} \tag{4.11}$$

Substituting (4.11) into (4.10), we obtain

$$\begin{aligned}
\lim_{s_1\to\infty} f_{GF}(t) &= \frac{1}{t\sigma\Gamma(s_2)} \lim_{s_1\to\infty} (s_1+s_2)^{s_2} \left(\frac{\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)}{1+\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)}\right)^{s_1} \left(1+\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)\right)^{-s_2} \\
&= \frac{1}{t\sigma\Gamma(s_2)} \lim_{s_1\to\infty} \left(\frac{s_1+s_2}{1+\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)}\right)^{s_2} \lim_{s_1\to\infty} \left(\frac{\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)}{1+\frac{s_1}{s_2}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)}\right)^{s_1} \\
&= \frac{1}{t\sigma\Gamma(s_2)} \left(s_2\exp\left(-\frac{\log(t)-\mu}{\sigma}\right)\right)^{s_2} \lim_{s_1\to\infty} \left(1+\frac{s_2\exp\left(-\frac{\log(t)-\mu}{\sigma}\right)}{s_1}\right)^{-s_1} \\
&= \frac{s_2^{s_2}}{t\sigma\Gamma(s_2)} \exp\left(-s_2\frac{\log(t)-\mu}{\sigma}\right) \exp\left(-s_2\exp\left(-\frac{\log(t)-\mu}{\sigma}\right)\right). \qquad (4.12)
\end{aligned}$$

Now we let $\sigma_e = \sigma/\sqrt{s_2}$, so (4.12) becomes

$$\begin{aligned}
\lim_{s_1\to\infty} f_{GF}(t) &= \frac{s_2^{s_2-1/2}}{t\sigma_e\Gamma(s_2)} \exp\left(-s_2^{1-1/2}\frac{\log(t)-\mu}{\sigma_e}\right) \exp\left(-s_2\exp\left(-s_2^{-1/2}\frac{\log(t)-\mu}{\sigma_e}\right)\right) \\
&= \frac{s_2^{s_2-1/2}}{t\sigma_e\Gamma(s_2)} \exp\left(-s_2^{1/2}\frac{\log(t)-\mu}{\sigma_e} - s_2\exp\left(-s_2^{-1/2}\frac{\log(t)-\mu}{\sigma_e}\right)\right) \\
&= \frac{s_2^{s_2-1/2}}{t\sigma_e\Gamma(s_2)} \exp\left(-s_2^{1/2}\frac{\log(t)-\mu}{\sigma_e} - s_2\exp\left(-s_2^{-1/2}\frac{\log(t)-\mu}{\sigma_e}\right)\right). \qquad (4.13)
\end{aligned}$$

Letting $\mu_e = \mu$ and $Q = -s_2^{-1/2}$, it is clear that the random variable T^* follows an extended generalised gamma distribution, whose cumulative distribution function is defined by (1.5).

While Peng, Dear and Denham (1998) noted that the extended generalised gamma distribution is a special case of the generalised F distribution, we elaborate on this point and conclude that when $s_1 \to \infty$, the generalised F distribution reduces to the extended generalised gamma distribution with $Q < 0$ since s_2 is strictly positive; when $s_2 \to \infty$, it reduces to the extended generalised gamma distribution with $Q > 0$; when both s_1 and s_2 approach infinity (in either order), it reduces to the log-normal distribution.

4.2.3 The generalised logistic family

Consider the location and scale model

$$\log T^* = \mu + \sigma W,$$

where W has a logistic distribution with density

$$f_l(w) = \frac{e^w}{(1+e^w)^2}, \quad -\infty < w < \infty.$$

The variable T^* then follows a log-logistic distribution with density

$$\begin{aligned} f_{ll}(t) &= f_l\left(w = \frac{\log t - \mu}{\sigma}\right)\left|\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\log t - \mu}{\sigma}\right)\right| \\ &= \frac{1}{t\sigma}\exp\left(\frac{\log t - \mu}{\sigma}\right)\left(1+\exp\left(\frac{\log t - \mu}{\sigma}\right)\right)^{-2}, \quad -\infty < \mu < \infty,\ t, \sigma > 0. \end{aligned} \tag{4.14}$$

The log-logistic distribution is a commonly used lifetime distribution that is not contained in the generalised gamma family. It is useful because it has a similar shape to the log-normal distribution but with a heavier tail and is capable of modelling mortality rates that increase initially but decrease later. At the same time, its cumulative distribution function has an explicit form.

There are several different generalisations of the logistic distribution in the literature. Among them, Johnson et al. (1995) summarised four forms that are used by researchers. Before introducing them, we note that they can all be expanded to include location and scale parameters.

The Type I generalised logistic distribution has a probability density function defined as

$$f_{gl1}(x; a) = \frac{ae^{-x}}{(1+e^{-x})^{a+1}}, \quad -\infty < x < \infty,\ a > 0. \tag{4.15}$$

The skewing parameter a controls its skewness. For example, the distribution is negatively skewed for $0 < a < 1$ and positively skewed for $a > 1$. Nevertheless, its probability density function is always unimodal and log-concave. With X being a random variable with the Type I generalised logistic distribution, Ahuja and Nash (1967) noted that $-aX$ behaves like a standard exponential random variable when a is close to 0, and $X - \log(a)$ behaves like an extreme value random variable. In the context of survival analysis, Gupta and Kundu (2010) noted that the Type I generalised logistic distribution can be regarded as a family of proportional reverse hazards distribution functions with the logistic distribution as the baseline distribution. The proportional reversed hazards models are similar to the proportional hazards models, but instead of assuming the hazard functions are proportional, it is assumed that the reversed hazard functions, defined by $f(t)/F(t)$, are proportional.

If a random variable X follows the Type I generalised logistic distribution, then $-X$ follows a Type II generalised logistic distribution, whose probability density function has the form

$$\begin{aligned} f_{gl2}(x;a) &= f_{gl1}(-x;a)|-1| \\ &= \frac{ae^x}{(1+e^x)^{a+1}} \\ &= ae^x\left(\frac{e^{-x}+1}{e^{-x}}\right)^{-(a+1)} \\ &= ae^x\frac{e^{-ax}e^{-x}}{(e^{-x}+1)^{a+1}} \\ &= \frac{ae^{-ax}}{(1+e^{-x})^{a+1}}, \quad -\infty < x < \infty,\ a > 0. \end{aligned}$$

In survival analysis, among all four types of generalisations, most researchers prefer to work with either Type I or II generalisations of the logistic distribution.

The Type III generalisation has a density

$$f_{gl3}(x;s) = \frac{\Gamma(s+s)}{\Gamma(s)^2}\frac{e^{-sx}}{(1+e^{-x})^{2s}}, \quad -\infty < x < \infty,\ s > 0, \tag{4.16}$$

and we note that

$$\begin{aligned} f_{gl3}(-x;s) &= \frac{\Gamma(s+s)}{\Gamma(s)^2}\frac{e^{sx}}{(1+e^x)^{2s}} \\ &= \frac{\Gamma(s+s)}{\Gamma(s)^2}\left(\frac{e^x}{e^{2x}+2e^x+1}\right)^s \\ &= \frac{\Gamma(s+s)}{\Gamma(s)^2}\left(\frac{e^{x-2x}}{e^{2x-2x}+2e^{x-2x}+e^{-2x}}\right)^s \\ &= \frac{\Gamma(s+s)}{\Gamma(s)^2}\left(\frac{e^{-x}}{(1+e^{-x})^2}\right)^s \\ &= \frac{\Gamma(s+s)}{\Gamma(s)^2}\left(\frac{e^{-x}}{(1+e^{-x})^2}\right)^s \\ &= f_{gl3}(x;s). \end{aligned}$$

Hence, the probability density function (4.16) is an even function, so the Type III generalised logistic distribution is symmetric about 0 for every s. It has a thicker tail than the normal distribution, and as noted by Gumbel (1944), it has a close relationship with extreme value random variables.

The Type IV generalised logistic distribution contains the other three generalisations as special cases. It is well-studied by Johnson et al. (1995) and Nassar and Elmasry (2012). Johnson et al. (1995) gave its probability density function as

$$f_{gl}(x) = \frac{\Gamma(s_1+s_2)}{\Gamma(s_1)\Gamma(s_2)} \frac{e^{-s_1 x}}{(1+e^{-x})^{s_1+s_2}}, \quad -\infty < x < \infty,\ s_1, s_2 > 0, \tag{4.17}$$

with s_1 and $s_2 > 0$. The Type IV generalisation is sometimes referred to as the log F model, and it is equivalent to the two-parameter generalised F model after transforming the variable. To show this, recall the probability density function for $\epsilon = \frac{\log(T^*)-\mu}{\sigma}$ given by (4.2) was

$$f_\varepsilon(\varepsilon; s_1, s_2) = \frac{\Gamma(s_1+s_2)}{\Gamma(s_1)\Gamma(s_2)} \left(\frac{s_1}{s_2}\right)^{s_1} \exp(\varepsilon)^{s_1} \left(1 + \frac{s_1}{s_2}\exp(\varepsilon)\right)^{-s_1-s_2}.$$

Consider the transformation $\varepsilon = -X - \log(s_1/s_2)$. Then the density function for X is

$$\begin{aligned}
f_X(x) &= f_\varepsilon(\varepsilon = -x - \log(s_1/s_2)) \left|\frac{\mathrm{d}}{\mathrm{d}x}\left(-x - \log(s_1/s_2)\right)\right| \\
&= \frac{\Gamma(s_1+s_2)}{\Gamma(s_1)\Gamma(s_2)} \left(\frac{s_1}{s_2}\right)^{s_1} \frac{\exp\left(-s_1 x - \log\left((s_1/s_2)^{s_1}\right)\right)}{\left(1 + \frac{s_1}{s_2}\exp(-x - \log(s_1/s_2))\right)^{s_1+s_2}} \\
&= \frac{\Gamma(s_1+s_2)}{\Gamma(s_1)\Gamma(s_2)} \left(\frac{s_1}{s_2}\right)^{s_1} \frac{\exp\left(-s_1 x\right) / \left(s_1/s_2\right)^{s_1}}{\left(1 + \frac{s_1}{s_2}\exp\left(-x\right) / \left(s_1/s_2\right)\right)^{s_1+s_2}} \\
&= \frac{\Gamma(s_1+s_2)}{\Gamma(s_1)\Gamma(s_2)} \frac{\exp\left(-s_1 x\right)}{\left(1+\exp\left(-x\right)\right)^{s_1+s_2}},
\end{aligned}$$

which is (4.17).

To apply the Type IV generalised logistic distribution in the context of cure models, we need to first write the random variable X as a function of the susceptibles' survival times, T^*

$$\begin{aligned}
\varepsilon = \frac{\log T^* - \mu}{\sigma} &= -X - \log(s_1) + \log(s_2) \\
-X &= \frac{\log T^* - (\mu - \sigma\log(s_1) + \sigma\log(s_2))}{\sigma} \\
-X &:= \frac{\log T^* - \mu_1}{\sigma_1}.
\end{aligned} \tag{4.18}$$

Therefore, we note that a generalised F distribution with parameters s_1, s_2, μ and σ is the same as the distribution of a random variable $-X$, where X follows a Type IV generalised log-logistic distribution with parameters $s_1, s_2, \mu_1 = \mu - \sigma\log(s_1) + \sigma\log(s_2)$ and $\sigma_1 = \sigma$.

When $s_1 = s_2 = s$, the Type IV generalisation (4.17) reduces to the Type III generalisation

(4.16). In that case, since we showed that (4.16) is an even function, $X = -\frac{\log T^* - \mu}{\sigma}$ follows the same Type III generalised logistic distribution as the random variable $-X = \frac{\log T^* - \mu}{\sigma}$. In other words, a generalised F distribution with parameters μ, σ, s_1, s_2 reduces to the Type III generalised log-logistic distribution with parameters $\mu_1 = \mu, \sigma_1 = \sigma$ and s when $s_1 = s_2 = s$, whose density function is

$$\begin{aligned} f_{gll3}(t;\mu,\sigma,s) &= \frac{\Gamma(2s)}{t\sigma\Gamma(s)^2}\exp\left(s\frac{\log(t)-\mu}{\sigma}\right)\left(1+\exp\left(\frac{\log(t)-\mu}{\sigma}\right)\right)^{-2s} \\ &= \frac{\Gamma(2s)}{t\sigma\Gamma(s)^2}\exp\left(-s\frac{\log(t)-\mu}{\sigma}\right)\left(1+\exp\left(-\frac{\log(t)-\mu}{\sigma}\right)\right)^{-2s}, \end{aligned} \tag{4.19}$$

where $t, \sigma, s > 0$ and $-\infty < \mu < \infty$. Furthermore, when $s = 1$, (4.19) reduces to the standard log-logistic distribution, where the probability density function becomes

$$\begin{aligned} f_{ll}(t) &= \frac{1}{t\sigma}\exp\left(-\frac{\log(t)-\mu}{\sigma}\right)\left(1+\exp\left(-\frac{\log(t)-\mu}{\sigma}\right)\right)^{-2} \\ &= \frac{1}{t\sigma}\exp\left(\frac{\log(t)-\mu}{\sigma}\right)\left(1+\exp\left(\frac{\log(t)-\mu}{\sigma}\right)\right)^{-2}, \quad -\infty < \mu < \infty,\ t, \sigma > 0, \end{aligned}$$

the same as (4.14).

Now for a random variable X following the Type IV generalised logistic distribution, if we only fix $s_1 = 1$ without fixing s_2, then the probability density function for X is

$$\begin{aligned} f_{gl}(x|s_1 = 1) &= \frac{\Gamma(1+s_2)}{\Gamma(s_2)}\frac{e^{-x}}{(1+e^{-x})^{1+s_2}} \\ &= \frac{s_2 e^{-x}}{(1+e^{-x})^{1+s_2}}, \quad -\infty < x < \infty,\ s_2 > 0. \end{aligned}$$

The above is equivalent to the probability density function of the Type I generalised logistic distribution (4.15) with $a = s_2$. In other words, when $s_1 = 1$, $-X$ becomes a Type II generalised logistic distribution. That is, referring to the relationship (4.18), when $s_1 = 1$, a random variable T^* that follows the generalised F distribution with parameters s_1, s_2, μ and σ reduces to the Type II generalised log-logistic distribution with parameters $a = s_2, \mu_1 = \mu + \sigma\log(s_2)$ and $\sigma_1 = \sigma$,

whose probability density function is

$$\begin{aligned} f_{gll2}(t) &= f_{gl}(x = -\frac{\log t - \mu_1}{\sigma_1} | s_1 = 1, s_2 = a) \left| \frac{\mathrm{d}}{\mathrm{d}t} \left(-\frac{\log t - \mu_1}{\sigma_1} \right) \right| \\ &= \frac{a}{t\sigma} \exp\left(\frac{\log t - \mu_1}{\sigma_1} \right) \left(1 + \exp\left(\frac{\log t - \mu_1}{\sigma_1} \right) \right)^{-1-a} \\ &= \frac{1}{t\sigma} s_2 \exp\left(\frac{\log t - \mu - \sigma \log(s_2)}{\sigma} \right) \left(1 + \exp\left(\frac{\log t - \mu - \sigma \log(s_2)}{\sigma} \right) \right)^{-1-s_2} \\ &= \frac{1}{t\sigma} \exp\left(\frac{\log t - \mu}{\sigma} \right) \left(1 + \frac{1}{s_2} \exp\left(\frac{\log t - \mu}{\sigma} \right) \right)^{-1-s_2} \\ &= f_{GF}(t; s_1 = 1). \end{aligned} \tag{4.20}$$

If we further let $s_2 \to \infty$, then (4.20) reduces to the Weibull distribution with parameterisation as discussed in Section 4.2.2, as we would expect.

It is also interesting to note that the Gumbel and Gompertz distributions are closely related to the generalised log-logistic family of distributions. In this case, define $X_1 = Y - \log(a)$, with Y following a Type I generalised logistic distribution with parameter $a > 0$. When $a \to \infty$, the probability density function for X_1 is

$$\begin{aligned} \lim_{a\to\infty} f_{X_1}(x_1) &= \lim_{a\to\infty} f_{gl1}(y = x_1 + \log(a)) \left| \frac{\mathrm{d}}{\mathrm{d}x_1} (x_1 + \log(a)) \right| \\ &= \lim_{a\to\infty} \frac{a e^{-x_1 - \log(a)}}{\left(1 + e^{-x_1 - \log(a)}\right)^{1+a}} \\ &= \lim_{a\to\infty} \frac{e^{-x_1}}{\left(1 + e^{-x_1}/a\right)^{1+a}} \\ &= e^{-x_1} \lim_{a\to\infty} \left(1 + \frac{e^{-x_1}}{a} \right)^{-1} \lim_{a\to\infty} \left[\left(1 + \frac{e^{-x_1}}{a} \right)^{a} \right]^{-1} \\ &= e^{-x_1} e^{-e^{x_1}} \\ &= \exp(-x_1 - \exp(-x_1)), \end{aligned} \tag{4.21}$$

using (4.2.2) again. The function (4.21) above can be recognized as the kernel for the standard Gumbel distribution, which can be expanded using a location and scale transformation $X_2 =$

$(X_1 + \eta + \log(\eta))\, b^{-1}$, with $\eta, b > 0$, so

$$\begin{aligned} f_{gom}(x_2) &= \lim_{a\to\infty} f_{X_1}(x_1 = -bx_2 - \eta - \log(\eta)) \left| \frac{\mathrm{d}}{\mathrm{d}x_2} \left(-bx_2 - \eta - \log(\eta)\right)\right| \\ &= \exp(bx_2 + \eta + \log(\eta) - \exp(+bx_2 + \eta + \log(\eta)))b \\ &= b\eta \exp(bx_2 + \eta - \eta \exp(+bx_2 + \eta)). \end{aligned}$$

This is the probability density function of a Gompertz distribution, another parametric lifetime distribution that attracts much attention in survival analysis, especially in modelling mortality rates. The Gompertz distribution is not contained as a sub-model in the generalised F distribution.

Specifically, recall that if $\varepsilon = \frac{\log(T^*)-\mu}{\sigma}$ follows a two-parameter generalised F distribution with probability density function (4.2), the random variable $X = -\varepsilon - \log(s_1/s_2) = -\varepsilon - \log(s_1) + \log(s_2)$ follows a two-parameter Type IV generalised logistic distribution. This random variable X reduces to the Type I generalised logistic distribution with parameter a when $s_1 = 1$ and $s_2 = a$, that is, $X = -\varepsilon + \log(a)$ has the same distribution as Y. Hence, we write

$$\begin{aligned} X_2 &= \frac{X_1 + \eta + \log(\eta)}{b} \\ &= \frac{-\varepsilon + \log(a) - \log(a) + \eta + \log(\eta)}{b} \\ &= -\frac{\varepsilon - (\eta + \log(\eta))}{b} \\ &= -\frac{\log T^* - (\mu + \sigma\eta + \sigma\log(\eta))}{b\sigma} \\ &:= -\frac{\log T^* - \mu_2}{\sigma_2} \\ &= \frac{\log(1/T^*) + \mu_2}{\sigma_2}. \end{aligned}$$

That is, if T^* follows a generalised F distribution with parameters s_1, s_2, μ and σ, then $X_2 = \frac{\log(1/T^*)+\mu_2}{\sigma_2}$, a transformation of the random variable $1/T^*$, follows a Gompertz distribution with parameters η and b when $s_1 = 1$ and $s_2 \to \infty$.

4.2.4 The Burr distribution

Burr (1942) introduced many forms of cumulative distribution functions that can be fitted to data. Among them, the Type XII Burr family of distributions attracted the most attention. Its

cumulative distribution function is given as

$$F_b(y) = 1 - (1 + y^c)^{-k}, \tag{4.22}$$

with both c and k strictly positive and $y > 0$. The corresponding probability density function obtained by differentiating (4.22) is

$$f_b(y) = kc\frac{y^{c-1}}{(1 + y^c)^{k+1}}. \tag{4.23}$$

Of course, the probability density function (4.23) can be extended by adding location and scale parameters.

Consider a random variable ε following a two-parameter generalised F distribution with probability density function (4.2). Let $s_1 = 1$, $s_2 = k$ and then make the transformation

$$Y = \left(\frac{\exp(\varepsilon)}{k}\right)^{\frac{1}{c}},$$

or equivalently,

$$\varepsilon = \log k + c \log Y, \quad \exp(\varepsilon) = kY^c. \tag{4.24}$$

The probability density function of Y is

$$\begin{aligned} f_Y(y) &= f_\varepsilon\left(\varepsilon = \log k + c\log y | s_1 = 1, s_2 = k\right)\left|\frac{\mathrm{d}}{\mathrm{d}y}\left(\log k + c\log y\right)\right| \\ &= \frac{\Gamma(1+k)}{\Gamma(1)\Gamma(k)}\left(\frac{1}{k}\right)^1 (ky^c)\left(1 + \frac{1}{k}ky^c\right)^{-1-k}\frac{c}{y} \\ &= ky^c\left(1 + y^c\right)^{-1-k}\frac{c}{y} \\ &= kc\frac{y^{c-1}}{(1 + y^c)^{k+1}}, \end{aligned}$$

which is (4.23). This transformation (4.24) implies that, for T^* following a four-parameter generalised F distribution and Y a Type XII Burr distribution, we can write

$$\begin{aligned}
\log T^* &= \mu + \sigma\varepsilon \\
\log T^* &= \mu + \sigma \log k + c\sigma \log Y \\
\log Y &= \frac{\log T^* - \mu - \sigma \log k}{c\sigma} \\
Y &:= \exp\left(\frac{\log T^* - \mu_b}{\sigma_b}\right),
\end{aligned}$$

where $\mu_b = \mu + \sigma \log k$ and $\sigma_b = c\sigma$. We can also interpret the relationship between T^* and Y as

$$\log T^* = \mu_b + \sigma_b \log Y,$$

where Y has the density of a Type XII Burr distribution. In this case, the probability density function for T^* is

$$\begin{aligned}
f_b(t) &= f_b\left(y = \exp\left(\frac{\log t - \mu_b}{\sigma_b}\right)\right)\left|\frac{\mathrm{d}}{\mathrm{d}t}\left(\exp\left(\frac{\log t - \mu_b}{\sigma_b}\right)\right)\right| \\
&= kc\frac{y^{c-1}}{(1+y^c)^{k+1}}\frac{y}{t\sigma_b} \\
&= \frac{kc}{t\sigma_b}\exp\left(\frac{\log t - \mu_b}{c^{-1}\sigma_b}\right)\left(1+\exp\left(\frac{\log t - \mu_b}{c^{-1}\sigma_b}\right)\right)^{-k-1} \\
&= \frac{k}{t\sigma}\exp\left(\frac{\log t - \mu - \sigma\log k}{\sigma}\right)\left(1+\exp\left(\frac{\log t - \mu - \sigma\log k}{\sigma}\right)\right)^{-k-1} \\
&= \frac{1}{t\sigma}\exp\left(\frac{\log t - \mu}{\sigma}\right)\left(1+\frac{1}{k}\exp\left(\frac{\log t - \mu}{\sigma}\right)\right)^{-k-1},
\end{aligned}$$

which is equivalent to the probability density function of the Type II generalised log-logistic distribution (4.20) with $k = s_2$. Therefore, although in the literature, most authors claim the generalised F distribution reduces to the Type XII Burr distribution when $s_1 = 1$, it reduces to the log of a Type XII Burr distribution.

The Burr Type III distribution is contained in the generalised F family as well. The Burr Type III has a cumulative distribution function

$$F_{b3}(y) = (y^{-c}+1)^{-k},$$

again with $y, c, k > 0$. The corresponding density is

$$f_{b3}(t) = kc\frac{y^{-c-1}}{(y^{-c}+1)^{k+1}} = kc\frac{y^{-c-1}y^{ck+c}}{(y^{-c}+1)^{k+1}(y^c)^{k+1}} = kc\frac{y^{ck-1}}{(1+y^c)^{k+1}}. \tag{4.25}$$

Again consider a random variable ε following a generalised F distribution with $s_1 = k$ and $s_2 = 1$. Now make the transformation $\varepsilon = -\log k + c\log Y_1$, or equivalently, $\exp(\varepsilon) = Y_1^c/k$. The probability density function for Y_1 is

$$\begin{aligned} f_{B_1}(y) &= f_\varepsilon(\varepsilon = -\log k + c\log y | s_1 = k, s_2 = 1)\left|\frac{\mathrm{d}}{\mathrm{d}y}(-\log k + c\log y)\right| \\ &= \frac{\Gamma(k+1)}{\Gamma(k)\Gamma(1)}k^k\left(\frac{y^c}{k}\right)^k(1+ky^c/k)^{-k-1}\frac{c}{y} \\ &= k^{1+k}\frac{y^{ck}}{k^k}\left(1+y^{-c}\right)^{k-1}\frac{c}{y} \\ &= kc\frac{y^{ck-1}}{(1+y^c)^{k+1}}, \end{aligned}$$

which is (4.25). Similarly, the relationship with T^* and Y_1 is

$$\log T^* = \mu_{b3} + \sigma_{b3}\log Y_1,$$

where Y_1 is a Type III Burr random variable, $\mu_{b3} = \mu - \sigma\log k$ and $\sigma_{b3} = c\sigma$. We conclude that the generalised F distribution reduces to the log of a Type III Burr distribution when $s_2 = 1$.

4.2.5 Other distributions

In the previous subsection, we mentioned that the generalised gamma distribution reduces to the Weibull distribution when $\gamma = 1$, and hence the generalised F distribution reduces to the Weibull distribution when $s_1 = 1$ and $s_2 \to \infty$. On the other hand, when $s_1 \to \infty$, the generalised F distribution reduces to the extended generalised distribution with $Q < 0$, whose probability density function was derived as (4.13). If we then further take $s_2 = 1$, the distribution is,

however, not a Weibull distribution. We get

$$\begin{aligned}
\lim_{s_1\to\infty} f_{GF}(t;s_2=1) &= \frac{s_2^{s_2-1/2}}{t\sigma/\sqrt{s_2}\Gamma(s_2)}\exp\left(-s_2^{1/2}\frac{\log(t)-\mu}{\sigma/\sqrt{s_2}} - s_2\exp\left(-s_2^{-1/2}\frac{\log(t)-\mu}{\sigma/\sqrt{s_2}}\right)\right)\Big|_{s_2=1} \\
&= \frac{1}{t\sigma}\exp\left(-\frac{\log(t)-\mu}{\sigma} - \exp\left(-\frac{\log(t)-\mu}{\sigma}\right)\right) \\
&= \frac{1}{t\sigma}\left(t^{-1}\exp(\mu)\right)^{\sigma^{-1}}\exp\left(-\left(t^{-1}\exp(\mu)\right)^{\sigma^{-1}}\right) \\
&= t^{-(\sigma^{-1}+1)}\sigma^{-1}\left(\exp(\mu)\right)^{\sigma^{-1}}\exp\left(-\left(\frac{\exp(\mu)}{t}\right)^{\sigma^{-1}}\right) \\
&= t^{-(\alpha_w+1)}\alpha_w\lambda_w^{\alpha_w}\exp\left(-\left(\frac{\lambda_w}{t}\right)^{\alpha_w}\right), \qquad (4.26)
\end{aligned}$$

which is the probability density function of an inverse Weibull distribution with parameters $\alpha_w = \sigma^{-1}$ and $\lambda_w = \exp(\mu)$. The inverse Weibull distribution occurs as a limiting distribution of the largest order statistics.

The generalised Pareto distribution also plays an important role in extreme value theory. It has density

$$f(x) = \frac{1}{\sigma_p}\left(1+\frac{\xi(x-\mu_p)}{\sigma_p}\right)^{-\frac{1}{\xi}-1}, \quad \mu_p, \xi \in (-\infty,\infty), \sigma_p > 0, \qquad (4.27)$$

where $x \geq \mu_p$ if $\xi \geq 0$ and $\mu_p \leq x \leq \mu_p - \sigma_p/\xi$ if $\xi < 0$. We note that the generalised F model can be treated as a generalisation of the generalised Pareto distribution when $\xi > 0$. For ε following a two-parameter generalised F distribution with probability density function (4.2), let $s_1 = 1$ and $s_2 = \frac{1}{\xi}$, and consider the transformation $\varepsilon = \log\left(\frac{X-\mu_p}{\sigma_p}\right)$. The density function for X is

$$\begin{aligned}
f_X(x) &= f_\varepsilon(\varepsilon = \log\left(\frac{x-\mu_p}{\sigma_p}\right)|s_1=1, s_2=\xi^{-1})\left|\frac{\mathrm{d}}{\mathrm{d}x}\left(\log\left(\frac{x-\mu_p}{\sigma_p}\right)\right)\right| \\
&= \frac{\Gamma(1+\xi^{-1})}{\Gamma(1)\Gamma(\xi^{-1})}\xi\frac{x-\mu_p}{\sigma_p}\left(1+\xi\frac{x-\mu_p}{\sigma_p}\right)^{-1-\frac{1}{\xi}}\frac{1}{x-\mu_p} \\
&= \frac{1}{\sigma_p}\left(1+\frac{\xi(x-\mu_p)}{\sigma_p}\right)^{-\frac{1}{\xi}-1}, \qquad (4.28)
\end{aligned}$$

which is equivalent to (4.27) when $\xi > 0$. However, the generalised F model cannot contain the cases where $\xi \leq 0$ since s_2 is strictly positive, except when $\xi = \mu_p = 0$, which is a limiting case equivalent to the exponential distribution.

4.2.6 Summary

A summary table of the relationships between the generalised F distribution and some of its sub-models is provided in Table 4.1.

Lifetime Distribution	s_1	s_2	μ	σ
Generalised F (s_1, s_2, μ, σ)	s_1	s_2	μ	σ
Generalised Gamma $(\alpha, \lambda, \gamma)$	γ	∞	$\alpha^{-1}\log\gamma - \log\lambda$	α^{-1}
Weibull (α, λ)	1	∞	$-\log\lambda$	α^{-1}
Exponential (λ)	1	∞	$-\log\lambda$	1
Rayleigh (λ)	1	∞	$-\log\lambda$	0.5
Gamma (λ, γ)	γ	∞	$\log\gamma - \log\lambda$	1
Extended generalised Gamma $(Q > 0, \mu_e, \sigma_e)$	Q^{-2}	∞	μ_e	$\sigma_e Q^{-1}$
Log-Normal (μ_e, σ_e)	∞	∞	μ_e	$\sigma_e\sqrt{s_1}$
Extended generalised Gamma $(Q < 0, \mu_e, \sigma_e)$	∞	Q^{-2}	μ_e	$\sigma_e\lvert Q\rvert^{-1}$
Inverse Weibull (α_w, λ_w)	∞	1	$\log\lambda_w$	α_w^{-1}
Type III generalised Log-Logistic (s, μ, σ)	s	s	μ	σ
Type II generalised Log-Logistic (a, μ_1, σ_1)	1	a	$\mu_1 - \sigma_1\log(a)$	σ_1
Log-Logistic (μ, σ)	1	1	μ	σ
(Log) Type XII Burr (k, c, μ_b, σ_b)	1	k	$\mu_b - \sigma_b\log(k)$	$c\sigma_b$
(Log) Type III Burr $(k, c, \mu_{b3}, \sigma_{b3})$	k	1	$\mu_{b3} + \sigma_{b3}\log(k)$	$c\sigma_{b3}$

TABLE 4.1: Generalised F Model's Sub-Models

4.3 Reparameterisation of the generalised F distribution

As shown in Section 4.2, the generalised F model reduces to the generalised gamma distribution as $s_2 \to \infty$, and it further reduces to the log-normal distribution when $s_1 \to \infty$ as well. However, parameter estimation for the generalised F model can be complicated and unstable when the true parameters are on the boundary points, i.e. infinity. To solve the problem and regularise parameter estimation, Prentice (1975) reparameterised the generalised F model by replacing s_1 and s_2 with Q and P_F, with $P_F \geq 0$, $Q \in \mathbb{R}$, and

$$s_1 = 2\left(Q^2 + 2P_F + Qc_F\right)^{-1}, \quad s_2 = 2\left(Q^2 + 2P_F - Qc_F\right)^{-1},$$

where $c_F = \left(Q^2 + 2P_F\right)^{1/2}$. Equivalently,

$$Q = \left(\frac{1}{s_1} - \frac{1}{s_2}\right)\left(\frac{1}{s_1} + \frac{1}{s_2}\right)^{-1/2}, \quad P_F = \frac{2}{s_1 + s_2}.$$

Recall that if $\varepsilon = \frac{\log T^* - \mu}{\sigma}$ follows a two-parameter generalised F distribution with parameters s_1 and s_2, then T^* has its probability density function in the form (4.3) with parameters s_1, s_2, μ, σ. After the reparameterisation, the convention is to redefine the two-parameter generalised F random variable ε as

$$\varepsilon = \frac{\log T^* - \mu}{c_F^{-1}\sigma},$$

and hence

$$\log T^* = \mu + c_F^{-1}\sigma\varepsilon := \mu_e + \sigma_e,$$

where $\mu_e = \mu$ and $\sigma_e = \left(Q^2 + 2P_F\right)^{-1/2}\sigma$. This further reparameterisation of the scale parameter is consistent with the parameters in an extended generalised gamma distribution. We now construct the relationship between the sub-models of the generalised F distribution and this reparameterised generalised F distribution and present the results in Table 4.2.

Lifetime Distribution	Q	P_F	μ_e	σ_e
Generalised F (Q, P, μ, σ)	Q	P_F	μ_e	σ_e
Extended generalised Gamma (Q, μ_e, σ_e)	Q	0	μ_e	σ_e
Weibull (α, λ)	1	0	$-\log\lambda$	α^{-1}
Exponential (λ)	1	0	$-\log\lambda$	1
Rayleigh (λ)	1	0	$-\log\lambda$	0.5
Gamma (λ, γ)	$\gamma^{-1/2}$	0	$\log\gamma - \log\lambda$	$\gamma^{1/2}$
Log-Normal (μ_e, σ_e)	0	0	μ_e	σ_e
Inverse Weibull (α_w, λ_w)	-1	0	$\log\lambda_w$	α_w^{-1}
Type III generalised Log-Logistic (s, μ, σ)	0	s^{-1}	μ	$\sqrt{\frac{\sigma^2 s}{2}}$
Type II generalised Log-Logistic (a, μ_1, σ_1)	$\frac{a-1}{\sqrt{a^2+a}}$	$\frac{2}{a+1}$	$\mu_1 - \sigma_1\log(a)$	$\frac{a\sigma_1}{a+1}$
Log-Logistic (μ, σ)	0	1	μ	$\frac{\sigma}{\sqrt{2}}$
(Log) Type XII Burr (k, c, μ_b, σ_b)	$\frac{k-1}{\sqrt{k^2+k}}$	$\frac{2}{k+1}$	$\mu_b - \sigma_b\log(k)$	$\frac{ck\sigma_b}{k+1}$
(Log) Type III Burr $(k, c, \mu_{b3}, \sigma_{b3})$	$\frac{1-k}{\sqrt{k^2+k}}$	$\frac{2}{k+1}$	$\mu_{b3} + \sigma_{b3}\log(k)$	$\frac{ck\sigma_{b3}}{k+1}$

TABLE 4.2: Reparameterised Generalised F Model's Sub-Models

4.4 Maximum Domain of Attraction

In Chapter 6, we will use extrapolation techniques from extreme value theory to estimate cure proportions under insufficient follow-up. Therefore, the maximum domain of attraction that the generalised F distribution and its sub-models belong to, if any, is of particular interest. Extreme value theory will be discussed in Section 6.3.1.

For a four-parameter generalised F distribution with probability density function (4.3), its corresponding survival function $S_{GF}(t) = \int_t^\infty f_{GF}(y)\mathrm{d}y$ is regularly varying (defined by Definition 6.1 in Section 6.3.1) at infinity with index $-s_2/\sigma < 0$. Thus the generalised F distribution is in the Fréchet domain of attraction. On the other hand, some of its sub-models satisfy von Mises' Condition ((4.29) in Proposition 4.4.1) and thus belong to the Gumbel maximum domain of attraction.

Proposition 4.4.1 (Proposition 1.18 in Resnick (1987))**.** Suppose F_0 has a finite negative second derivative F_0'' for all t in some left neighbourhood of τ_{F_0}. If

$$\lim_{t\uparrow\tau_{F_0}} \frac{F_0''(t)\,(1-F_0(t))}{(F_0'(t))^2} = -1, \tag{4.29}$$

then F_0 belongs to the Gumbel domain of attraction. Conversely, if F_0'' is non-decreasing, $F_0'(t) = \int_t^{\tau_{F_0}} (-F_0''(u))\,\mathrm{d}u$ and F_0 belongs to the Gumbel domain of attraction, then (4.29) holds.

In this section, we will show that the two-parameter generalised F distribution (also known as the log F distribution) and the generalised gamma distribution satisfy von Mises' condition.

4.4.1 The log F distribution in the Gumbel maximum domain of attraction

Recall (4.2), for $\varepsilon = \frac{\log(T^*)-\mu}{\sigma}$ following a log F distribution. Its density is

$$f_\varepsilon(\varepsilon; s_1, s_2) = \frac{\Gamma(s_1+s_2)}{\Gamma(s_1)\Gamma(s_2)} \left(\frac{s_1}{s_2}\right)^{s_1} \exp(\varepsilon)^{s_1} \left(1+\frac{s_1}{s_2}\exp(\varepsilon)\right)^{-s_1-s_2}.$$

It is then easy to see that the derivative of the density is

$$\begin{aligned}
f_\varepsilon'(\varepsilon) &= \frac{\Gamma(s_1+s_2)}{\Gamma(s_1)\Gamma(s_2)} \left(\frac{s_1}{s_2}\right)^{s_1} s_1 \exp(s_1\varepsilon) \left(1+\frac{s_1}{s_2}\exp(\varepsilon)\right)^{-s_1-s_2} + \\
&\quad \frac{\Gamma(s_1+s_2)}{\Gamma(s_1)\Gamma(s_2)} \left(\frac{s_1}{s_2}\right)^{s_1} \exp(s_1\varepsilon)\,(-s_1-s_2)\,\frac{s_1}{s_2}\exp(\varepsilon) \left(1+\frac{s_1}{s_2}\exp(\varepsilon)\right)^{-s_1-s_2-1} \\
&= \frac{\Gamma(s_1+s_2)}{\Gamma(s_1)\Gamma(s_2)} \left(\frac{s_1}{s_2}\right)^{s_1} \exp(s_1\varepsilon) \left(1+\frac{s_1}{s_2}\exp(\varepsilon)\right)^{-s_1-s_2} \times \\
&\qquad \left\{ s_1 - (s_1+s_2)\,\frac{s_1}{s_2}\exp(\varepsilon) \left(1+\frac{s_1}{s_2}\exp(\varepsilon)\right)^{-1} \right\} \\
&= f_\varepsilon(\varepsilon) \left\{ s_1 + (-s_1-s_2)\,\frac{s_1}{s_2}\exp(\varepsilon) \left(1+\frac{s_1}{s_2}\exp(\varepsilon)\right)^{-1} \right\},
\end{aligned} \tag{4.30}$$

and the survival function is

$$\begin{aligned}
1-F_\varepsilon(\varepsilon; s_1, s_2) &= \int_\varepsilon^\infty \frac{\Gamma(s_1+s_2)}{\Gamma(s_1)\Gamma(s_2)} \left(\frac{s_1}{s_2}\right)^{s_1} \exp(s_1 u) \left(1+\frac{s_1}{s_2}\exp(u)\right)^{-s_1-s_2} \mathrm{d}u \\
&= \frac{\Gamma(s_1+s_2)}{\Gamma(s_1)\Gamma(s_2)} \left(\frac{s_1}{s_2}\right)^{s_1} \int_\varepsilon^\infty \exp(s_1 u) \left(1+\frac{s_1}{s_2}\exp(u)\right)^{-s_1-s_2} \mathrm{d}u.
\end{aligned} \tag{4.31}$$

Substituting (4.30) and (4.31) into (4.29), we have

$$\begin{aligned}
\frac{f'_\varepsilon(\varepsilon)\,(1-F_\varepsilon(\varepsilon))}{(f_\varepsilon(\varepsilon))^2} &= \frac{f_\varepsilon(\varepsilon)\left\{s_1+(-s_1-s_2)\,\frac{s_1}{s_2}\exp(\varepsilon)\left(1+\frac{s_1}{s_2}\exp(\varepsilon)\right)^{-1}\right\}}{f_\varepsilon(\varepsilon)}\times \\
&\qquad \frac{\frac{\Gamma(s_1+s_2)}{\Gamma(s_1)\Gamma(s_2)}\left(\frac{s_1}{s_2}\right)^{s_1}\int_\varepsilon^\infty \exp(s_1u)\left(1+\frac{s_1}{s_2}\exp(u)\right)^{-s_1-s_2}\mathrm{d}u}{\frac{\Gamma(s_1+s_2)}{\Gamma(s_1)\Gamma(s_2)}\left(\frac{s_1}{s_2}\right)^{s_1}\exp(s_1\varepsilon)\left(1+\frac{s_1}{s_2}\exp(\varepsilon)\right)^{-s_1-s_2}} \\
&= \frac{1}{\exp(s_1\varepsilon)\left(1+\frac{s_1}{s_2}\exp(\varepsilon)\right)^{-s_1-s_2}}\times\int_\varepsilon^\infty \exp(s_1u)\left(1+\frac{s_1}{s_2}\exp(u)\right)^{-s_1-s_2}\mathrm{d}u\times \\
&\qquad \left\{s_1+(-s_1-s_2)\,\frac{s_1}{s_2}\exp(\varepsilon)\left(1+\frac{s_1}{s_2}\exp(\varepsilon)\right)^{-1}\right\} \\
&= \int_\varepsilon^\infty \exp(s_1u)\left(1+\frac{s_1}{s_2}\exp(u)\right)^{-s_1-s_2}\mathrm{d}u \Bigg/ \\
&\qquad \frac{\exp(s_1\varepsilon)\left(1+\frac{s_1}{s_2}\exp(\varepsilon)\right)^{-s_1-s_2}}{s_1+(-s_1-s_2)\,\frac{s_1}{s_2}\exp(\varepsilon)\left(1+\frac{s_1}{s_2}\exp(\varepsilon)\right)^{-1}} \\
&:= g(\varepsilon)/h(\varepsilon).
\end{aligned}$$

Let $A(\varepsilon)=(-s_1-s_2)\,\frac{s_1}{s_2}\exp(\varepsilon)\left(1+\frac{s_1}{s_2}\exp(\varepsilon)\right)^{-1}$. We then find

$$\begin{aligned}
g'(\varepsilon) &= -\exp(s_1\varepsilon)\left(1+\frac{s_1}{s_2}\exp(\varepsilon)\right)^{-s_1-s_2}, \\
h'(\varepsilon) &= s_1\frac{-g'(\varepsilon)}{\{s_1+A(\varepsilon)\}}+A(\varepsilon)\frac{-g'(\varepsilon)}{\{s_1+A(\varepsilon)\}}-\left\{A(\varepsilon)-\frac{A(\varepsilon)^2}{-s_1-s_2}\right\}\frac{-g'(\varepsilon)}{\{s_1+A(\varepsilon)\}^2}.
\end{aligned}$$

Using L'Hopital's theorem, we get

$$\begin{aligned}
\lim_{\varepsilon\to\infty} g(\varepsilon)/h(\varepsilon) &= \lim_{\varepsilon\to\infty} g'(\varepsilon)/h'(\varepsilon) \\
&= \lim_{\varepsilon\to\infty} -1\Bigg/\left\{\frac{s_1}{\{s_1+A(\varepsilon)\}}+\frac{A(\varepsilon)}{\{s_1+A(\varepsilon)\}}-\left[A(\varepsilon)-\frac{A(\varepsilon)^2}{-s_1-s_2}\right]\frac{1}{\{s_1+A(\varepsilon)\}^2}\right\} \\
&= \lim_{\varepsilon\to\infty} -1\Bigg/\left\{\left[s_1^2+s_1A(\varepsilon)+s_1A(\varepsilon)+A(\varepsilon)^2-A(\varepsilon)+\frac{A(\varepsilon)^2}{-s_1-s_2}\right]\frac{1}{\{s_1+A(\varepsilon)\}^2}\right\} \\
&= \lim_{\varepsilon\to\infty} -\frac{\{s_1+A(\varepsilon)\}^2}{(s_1+A(\varepsilon))^2-A(\varepsilon)\left(1-\frac{A(\varepsilon)}{-s_1-s_2}\right)} \\
&= \lim_{\varepsilon\to\infty} -\frac{\{s_1+A(\varepsilon)\}^2}{(s_1+A(\varepsilon))^2-A(\varepsilon)\left(1-\frac{s_1}{s_2}\exp(\varepsilon)\left(1+\frac{s_1}{s_2}\exp(\varepsilon)\right)^{-1}\right)}.
\end{aligned} \tag{4.32}$$

Note that

$$\begin{aligned}\lim_{\varepsilon\to\infty}\frac{s_1}{s_2}\exp(\varepsilon)\left(1+\frac{s_1}{s_2}\exp(\varepsilon)\right)^{-1} &= \frac{s_1}{s_2}\lim_{\varepsilon\to\infty}\frac{\exp(\varepsilon)}{1+\frac{s_1}{s_2}\exp(\varepsilon)}\\ &= \frac{s_1}{s_2}\times\frac{s_2}{s_1}\\ &= 1. \end{aligned} \tag{4.33}$$

Substituting (4.33) into (4.32), we find

$$\begin{aligned}\lim_{\varepsilon\to\infty}\frac{f'_\varepsilon(\varepsilon)\left(1-F_\varepsilon(\varepsilon)\right)}{\left(f_\varepsilon(\varepsilon)\right)^2} &= -\lim_{\varepsilon\to\infty}\frac{\left\{s_1+A(\varepsilon)\right\}^2}{\left(s_1+A(\varepsilon)\right)^2-A(\varepsilon)\left(1-\frac{s_1}{s_2}\exp(\varepsilon)\left(1+\frac{s_1}{s_2}\exp(\varepsilon)\right)^{-1}\right)}\\ &= -\lim_{\varepsilon\to\infty}\frac{\left\{s_1+A(\varepsilon)\right\}^2}{\left(s_1+A(\varepsilon)\right)^2-A(\varepsilon)\left(1-1\right)}\\ &= -1, \end{aligned} \tag{4.34}$$

which satisfies (4.29). We also note that the von Mises condition (4.29) holds for all choices of s_1 and s_2, which produce proper distributions.

4.4.2 The generalised gamma distribution in the Gumbel maximum domain of attraction

Recall the density function for a generalised gamma distribution $f_{gg}(t)$ (1.1). Its derivative is

$$f'_{gg}(t) = \frac{\alpha}{\Gamma(\gamma)}t^{\alpha\gamma-2}\lambda^{\alpha\gamma}\exp\left(-(\lambda t)^\alpha\right)\left[\alpha\gamma-1-\alpha(\gamma t)^\alpha\right]$$

and its survival function $1-F_{gg}(t)$ is

$$1-F_{gg}(t) = \frac{\alpha}{\Gamma(\gamma)}\lambda^{\alpha\gamma}\int_t^\infty u^{\alpha\gamma-1}\exp\left(-(\lambda u)^\alpha\right)\mathrm{d}u.$$

Defining

$$\begin{aligned} D(t) &= \frac{(1-F_{gg}(t))f'_{gg}(t)}{f^2_{gg}(t)} = \frac{f'_{gg}(t)}{f_{gg}(t)} \times \frac{1-F_{gg}(t)}{f_{gg}(t)} \\ &= \frac{\alpha\gamma - 1 - \alpha(\lambda t)^\alpha}{t} \times \frac{\int_t^\infty u^{\alpha\gamma-1}\exp\big(-(\lambda u)^\alpha\big)\mathrm{d}u}{t^{\alpha\gamma-1}\exp\big(-(\lambda t)^\alpha\big)} \\ &= \int_t^\infty u^{\alpha\gamma-1}\exp\big(-(\lambda u)^\alpha\big)\mathrm{d}u \Big/ \frac{t^{\alpha\gamma}\exp\big(-(\lambda t)^\alpha\big)}{\alpha\gamma - 1 - \alpha(\lambda t)^\alpha} \\ &= g(t)/h(t), \end{aligned}$$

we can find

$$\begin{aligned} g'(t) &= \frac{\mathrm{d}}{\mathrm{d}t}g(t) = -t^{\alpha\gamma-1}\exp\big(-(\lambda t)^\alpha\big) \\ h'(t) &= \frac{\mathrm{d}}{\mathrm{d}t}h(t) = \frac{t^{\alpha\gamma-1}\exp\big(-(\lambda t)^\alpha\big)\big[\alpha^2(\lambda t)^{2\alpha} + (\alpha^2 - 2\alpha^2\gamma + \alpha)(\lambda t)^\alpha + (\alpha\gamma-1)\alpha\gamma\big]}{\alpha^2(\lambda t)^{2\alpha} + 2\alpha(1-\alpha\gamma)(\lambda t)^\alpha + (\alpha\gamma-1)^2}. \end{aligned}$$

Using L'Hopital's theorem we get

$$\begin{aligned} \lim_{t\to\infty} g(t)/h(t) &= \lim_{t\to\infty} g'(t)/h'(t) \\ &= \lim_{t\to\infty} -\frac{\alpha^2(\lambda t)^{2\alpha} + 2\alpha(1-\alpha\gamma)(\lambda t)^\alpha + (\alpha\gamma-1)^2}{\alpha^2(\lambda t)^{2\alpha} + (\alpha^2 - 2\alpha^2\gamma + \alpha)(\lambda t)^\alpha + (\alpha\gamma-1)\alpha\gamma} \\ &= -1. \end{aligned}$$

So the von Mises condition (4.29) holds for all choices of α, λ and γ which produce proper distributions.

4.4.3 Summary

In Table 4.3, we list many common lifetime distributions which can be obtained as sub-models by choices of s_1 and s_2 in the generalised F and generalised Gamma distributions, along with the maximum domain of attraction that they belong to.

Lifetime Distribution					DoA
generalised F (s_1, s_2, μ, σ)	s_1	s_2	μ	σ	F
Type II generalised Log-Logistic (a, μ_1, σ_1)	1	a	$\mu_1 - \sigma_1 \log(a)$	σ_1	F
Type III generalised Log-Logistic (s, μ, σ)	s	s	μ	σ	F
Log-Logistic (μ, σ)	1	1	μ	σ	F
Log F (s_1, s_2)	s_1	s_2			G
generalised Gamma $(\alpha, \lambda, \gamma)$	γ	∞	$\alpha^{-1} \log \gamma - \log \lambda$	α^{-1}	G
Weibull (α, λ)	1	∞	$-\log \lambda$	α^{-1}	G
Exponential (λ)	1	∞	$-\log \lambda$	1	G
Rayleigh (λ)	1	∞	$-\log \lambda$	0.5	G
Gamma (λ, γ)	γ	∞	$\log \gamma - \log \lambda$	1	G
Log-Normal (μ_e, σ_e)	∞	∞	μ_e	$\sigma_e \sqrt{s_1}$	G

TABLE 4.3: *Sub-Models of the Generalised F and generalised Gamma Models with their maximum domain of attraction (DoA).*

The table indicates that the log-logistic and Type II and III generalised log-logistic distributions are sub-models of the generalised F distribution for finite choices of s_1 and s_2, consequently are in the Fréchet domain of attraction.

The other distributions listed in Table 4.3, including the generalised Gamma and its sub-models, are sub-models of the generalised F distribution for infinite limiting values of s_1 or s_2. These all belong to the Gumbel maximum domain of attraction.

When both s_1 and s_2 approach infinity, the generalised Gamma distribution reduces to the log-normal distribution, which belongs to the Gumbel maximum domain of attraction (Resnick 1987, p.53).

We remark that there are lifetime distributions satisfying von Mises' condition that are not contained in either the log F or generalised F distributions. An example is the Gompertz distribution. Also, there are distributions such as the geometric which are in no domain of maximum attraction, see Resnick (1987, p.45).

Chapter 5

Parametric mixture cure model selection

5.1 Introduction

With the connections between the reparameterised generalised F distribution and its sub-models clearly outlined and summarised in Table 4.2, we are now able to perform parametric mixture cure model selection. The primary tool we use in this section is the likelihood ratio test because "a number of arguments suggest that frequentists should base inference on the likelihood ratio", as noted by Welsh (1996, p.g. 133). The likelihood ratio contains all the information in the data about the model, and tests based on it are often optimal according to the Neyman-Pearson Lemma (see Section 3.3 of Welsh (1996)). Moreover, Garthwaite, Jolliffe and Jones (2002, p.g. 85) noted that the likelihood ratio test is the uniformly most powerful test under some assumptions, and it is asymptotically efficient and unbiased. Apart from these nice properties, the likelihood ratio test statistics are often feasible to calculate and easy to interpret.

Assume a model $\mathcal{F}$ is the set of distributions with parameters $\underset{\sim}{\theta}$ defined on the parameter space Θ. For a set of independent and identically distributed data with density function f contained in $\mathcal{F}$, let $L(\underset{\sim}{\theta})$ denote the likelihood function. For $\omega \in \Theta$, for testing

$$H_0\colon \underset{\sim}{\theta} \in \omega \quad \text{versus} \quad H_1\colon \underset{\sim}{\theta} \in \Theta \setminus \omega,$$

the likelihood ratio statistic is defined as

$$\Lambda(\omega) = \frac{\sup_{\theta \in \omega} L(\underset{\sim}{\theta})}{\sup_{\theta \in \Theta \setminus \omega} L(\underset{\sim}{\theta})}.$$

In cases where only a scalar parameter $\theta_k \in \underset{\sim}{\theta}$ is of interest, for example, when testing

$$H_0\colon \theta_k = \theta_{k0} \quad \text{versus} \quad H_1\colon \theta_k \neq \theta_{k0},$$

where θ_k is an element in $\underset{\sim}{\theta}$ and θ_{k0} is an arbitrary constant, the likelihood ratio statistic reduces to

$$\Lambda(\theta_{k0}) = \frac{\sup_{\theta_k=\theta_{k0}} L(\underset{\sim}{\theta})}{\sup_{\Theta} L(\underset{\sim}{\theta})}. \tag{5.1}$$

It is obvious that a small value of $\Lambda(\theta_{k0})$ indicates strong evidence against the null hypothesis. In addition, Cox and Hinkley (1979, Section 9.3) noted that $-2\log(\Lambda(\theta_{k0}))$ often follows a chi-squared distribution with 1 degree of freedom approximately. Therefore, an approximate rejection region and/or p-value for the test are easily obtained. In this case, since $0 < \Lambda(\theta_{k0}) \leq 1$, we note that a large value of $-2\log(\Lambda(\theta_{k0}))$ indicates strong evidence against the null hypothesis. For the rest of this chapter, we will use $-2\log(\Lambda(\theta_{k0}))$ as the test statistic for the likelihood ratio tests and refer to it as the likelihood ratio test statistics.

In addition, it is important to note that the conclusions regarding the asymptotic distribution of the likelihood ratio test statistic depend on certain regularity conditions. For instance, Chen et al. (2020) demonstrated through simulation studies that when the null model lies on the boundary of the parameter space, the above results can be misleading for all sample sizes. Specifically, Zhou and Maller (1995) found that when F_0 follows an exponential distribution, the asymptotic distribution of the likelihood ratio statistic is a 50-50 mixture of a chi-squared distribution with one degree of freedom and a point mass at 0. Expanding on this, Vu, Maller, and Zhou (1998) generalized this result to distributions within the exponential family under specific conditions. However, since the generalized F distribution does not belong to the exponential family, their results are not applicable in this context. For this chapter, our focus is on demonstrating the convenience of using the generalized F distribution as a starting point for parametric cure model selection. Therefore, we will continue to use the 50-50 mixture of a chi-squared distribution with one degree of freedom and a point mass at 0 as the sampling distribution for the likelihood ratio test statistic when the null model is on the boundary of the parameter space. This approach appears more conservative compared to using a simple Chi-squared distribution with one degree of freedom, which was proved to be misleading. We left it as an area for future research to validate this approach or to develop a more tailored asymptotic distribution through simulation studies and theoretical explorations.

With likelihood ratio tests introduced above, we outline the steps needed to perform model selection for parametric mixture cure models using the generalised F distribution in Section 5.2. After an optimal cure model is selected, we need to check whether or not the chosen model is a good fit for the data using a goodness of fit test, whose details are outlined in Section 5.3. Then we demonstrate the application of parametric cure model selection to both simulated data and breast cancer dataset in Sections 5.4-5.5.

5.2 Methodology

Step 1: Presence of immunes

Using the likelihood ratio test for model selection requires a flexible single parametric model, which embeds many sub-models of interest. We will use the reparameterised generalised F distribution constructed in Section 4.4. Given survival data with potential long-term survivors, researchers may prefer to start with a test for the presence of immunes. However, we do not recommend directly fitting the data with a generalised F mixture cure model and conducting a likelihood ratio test since this five-parameter model can overfit on some occasions, making the test result unreliable. We instead suggest using the nonparametric method proposed by Maller and Zhou (1996, p.g. 76) to test

$$\begin{aligned} &H_0\text{: } p = 1 \quad \text{(i.e. there are no immunes)}, \\ &H_1\text{: } p < 1 \quad \text{(i.e. there are immunes)}, \end{aligned} \tag{5.2}$$

where the critical values can be found in Maller and Zhou (1996, p.g. 254 Table A.2). If the test rejects the null hypothesis, we will fit parametric mixture cure models to compute the likelihood ratios for the subsequent tests in the following steps. Otherwise, we will fit parametric models directly to the data. As previously mentioned, as the test for the presence of immunes encounters difficulties due to the issue of the null model lying on the boundary of the parameter space, we will employ the 50-50 mixture of a chi-squared distribution with one degree of freedom and a point mass at 0 for the likelihood ratio test statistic.

On the other hand, we note that researchers may carry out the model selection process assuming there are immunes in the population without performing a test for the presence of immunes. In this way, researchers may choose to conduct a likelihood ratio test for the presence

of immunes with the optimal mixture cure model selected to avoid the problem of over-fitting to some extent. However, we note that such a likelihood ratio test may fail to reject the null hypothesis when the follow-up time is insufficient.

Step 2: Extended generalised gamma versus generalised log-logistic

In Step 2, we will test the following sets of hypotheses simultaneously with the same significance level:

H_0: F_0 should be reduced to an extended generalised gamma model,

H_1: F_0 should not be reduced to an extended generalised gamma model;

and

H_0: F_0 should be reduced to a Type III generalised log-logistic model,

H_1: F_0 should not be reduced to a Type III generalised log-logistic model.

The former test is equivalent to testing

$$H_0 : P_F = 0 \text{ versus } H_1 : P_F > 0, \tag{5.3}$$

and the latter is equivalent to testing

$$H_0 : Q = 0 \text{ versus } H_1 : Q \neq 0, \tag{5.4}$$

where P_F and Q are the parameters of the reparameterised generalised F distribution (see Table 4.2). When computing the test statistics for the tests, if we conclude there is not enough evidence for the presence of immunes in the study in **Step 1**, we will fit a generalised F distribution with parameters Q, P_F, μ and σ (see Section 4.3) to the data to compute the likelihood ratio. Again we note that the test with null hypothesis $P_F = 0$ lies on the boundary of parameter space, therefore we will use the 50-50 mixture of a chi-squared distribution with one degree of freedom and a point mass at 0 for that specific test. The modelling fitting is done using the "flexsurvcure" package in R and a detailed demonstration of the use of this R program can be found in Section

5.4.1. On the other hand, if we conclude that there is evidence for long-term survivors, we will instead fit the generalised F mixture cure model to data to compute $\Lambda(\underset{\sim}{\theta})$.

According to the results of the above tests:

- If tests of both (5.3) and (5.4) reject the null hypothesis, we will stop the model selection process and conclude that F_0 should be specified as a generalised F distribution.
- If the test of (5.3) fails to reject the null hypothesis and the test of (5.4) rejects the null hypothesis, then we will continue to Step 3 a); if the test of (5.3) rejects the null hypothesis and the test of (5.4) fails to reject the null hypothesis, we will continue to Step 3 b).
- If tests of both (5.3) and (5.4) fail to reject the null hypothesis, we will compare the likelihood ratio test statistics for the two hypothesis tests, recalling that a high likelihood ratio indicates strong evidence against the null hypothesis. If the likelihood ratio test statistic for testing (5.3) is smaller than the one for (5.4), we conclude F_0 should be reduced to an extended generalised gamma model and continue to Step 3 a). Alternatively, if the likelihood ratio for testing (5.4) is smaller than that for testing (5.3), we conclude F_0 should be reduced to a Type III generalised log-logistic model and continue to Step 3 b).

Step 3 a): 2-parameter sub-models of extended generalised gamma

In Tables 1.1 and 4.2 we listed the sub-models contained in the extended generalised gamma family. Among them, the Weibull, gamma, log-normal and inverse Weibull distributions have two parameters. To test whether F_0 should be reduced from an extended gamma model with parameters Q, μ_e and σ_e to one of the 2-parameter sub-models of it, we perform the following

four tests:

$$\begin{aligned} &H_0\text{: } Q = 1 \quad \text{(i.e. reduce to Weibull)},\\ &H_1\text{: } Q \neq 1 \quad \text{(i.e. do not reduce to Weibull)}; \end{aligned} \tag{5.5}$$

$$\begin{aligned} &H_0\text{: } Q = -1 \quad \text{(i.e. reduce to inverse Weibull)},\\ &H_1\text{: } Q \neq -1 \quad \text{(i.e. do not reduce to inverse Weibull)}; \end{aligned} \tag{5.6}$$

$$\begin{aligned} &H_0\text{: } Q = 0 \quad \text{(i.e. reduce to log-normal)},\\ &H_1\text{: } Q \neq 0 \quad \text{(i.e. do not reduce to log-normal)}. \end{aligned} \tag{5.7}$$

When testing whether the extended generalised gamma model should be reduced to a gamma model, we need to change the parameterisation of the extended generalised gamma distribution from $\{Q, \mu_e, \sigma_e\}$ to $\{Q, \alpha, \lambda\}$, using the relationships given in Section 4.2, without affecting the value of the likelihood ratio. The test to be conducted would then be

$$\begin{aligned} &H_0\text{: } \alpha = 1 \quad \text{(i.e. reduce to Gamma)},\\ &H_1\text{: } \alpha \neq 1 \quad \text{(i.e. do not reduce to Gamma)}. \end{aligned} \tag{5.8}$$

Similar to Step 2, according to the test results:

- If all of the four tests performed in Step 3 a) reject the null hypothesis, we stop the model selection process and conclude that F_0 should be specified as an extended generalised gamma distribution.
- If only one of the four tests fails to reject the null hypothesis, we will use that test result as our conclusion for Step 3. For example, if only the test of (5.5) fails to reject the null hypothesis, we conclude that F_0 should be reduced to the Weibull.
- If more than one test fails to reject the null hypothesis, we will compare the likelihood ratio test statistics and conclude Step 3 using the test result with the smallest observed test statistic (i.e. the strongest support for the null hypothesis).

Lastly, if the conclusion for Step 3 is F_0 should be reduced to a Weibull distribution, we continue to Step 4 a); if F_0 should be reduced to a gamma distribution, we continue to Step 4 b).

Step 3 b): log-logistic versus generalised log-logistic

If in Step 2 we concluded F_0 should be reduced to a Type III generalised log-logistic distribution with parameters s, μ, σ, then in Step 3, we test whether F_0 should be further reduced to a log-logistic distribution. That is, we test

$$\begin{aligned} &H_0\text{: } s = 1 \quad \text{(i.e. reduce to log-logistic)}, \\ &H_1\text{: } s \neq 1 \quad \text{(i.e. do not reduce to log-logistic)}. \end{aligned} \tag{5.9}$$

If the test fails to reject the null hypothesis, we will stop the model selection process and specify F_0 as a 2-parameter log-logistic distribution. If the test rejects the null hypothesis, we will stop the model selection process and specify F_0 as a 3-parameter Type III generalised log-logistic distribution.

Step 4 a): 1-parameter sub-models of Weibull

The two 1-parameter lifetime distributions widely applied in practice are the exponential distribution and the Rayleigh distribution. They are both contained in the Weibull distribution as sub-models. If the result from Step 3 suggests F_0 should be reduced to a Weibull distribution with parameters $\lambda = \exp(-\mu_e)$ and $\alpha = \sigma_e^{-1}$, we perform the following two tests

$$\begin{aligned} &H_0\text{: } \alpha = 1 \quad \text{(i.e. reduce to exponential)}, \\ &H_1\text{: } \alpha \neq 1 \quad \text{(i.e. do not reduce to exponential)}; \end{aligned} \tag{5.10}$$

$$\begin{aligned} &H_0\text{: } \alpha = 2 \quad \text{(i.e. reduce to Rayleigh)}, \\ &H_1\text{: } \alpha \neq 2 \quad \text{(i.e. do not reduce to Rayleigh)}. \end{aligned} \tag{5.11}$$

Similar to before, we will end the model selection process and conclude that F_0 should be Weibull if both tests reject the null hypothesis. We will compare the two likelihood ratios and use the result from the test with a smaller likelihood ratio as the overall conclusion if at least one of the tests fails to reject the null hypothesis.

Step 4 b): exponential versus gamma

The exponential distribution is a sub-model for not only the Weibull model but also the gamma model. If F_0 was concluded to be gamma with parameters $\gamma = Q^{-2}$ and λ in Step 3, we will perform the following test and use its result as the conclusion for the model selection process.

$$\begin{aligned} &H_0\text{: } \gamma = 1 \quad \text{(i.e. reduce to exponential)}, \\ &H_1\text{: } \gamma \neq 1 \quad \text{(i.e. do not reduce to exponential)}. \end{aligned} \tag{5.12}$$

In the case where the true F_0 follows an exponential distribution, we expect **Step 3 a)** of the model selection process to either conclude F_0 should be reduced to a gamma distribution or conclude F_0 should be reduced to a Weibull distribution. In practice, this imposes little restriction as we expect the likelihood ratio test in both **Step 4 a)** and **Step 4 b)** not to reject the null hypothesis.

The complete model selection process introduced in this section involves performing a sequence of conditional likelihood ratio tests, where we assume the same significance level for all of them. However, when multiple hypothesis tests are performed on the same data, the probability of obtaining a false positive result (i.e. Type I error) increases. Ang (2019) noted that the most popular techniques to lower this false positive rate include the Bonferroni correction and the Benjamini-Hochberg procedure. However, since we cannot know ahead of time how many likelihood ratio tests will be conducted for the model selection process for each dataset, neither of these methods is fully compatible in practice. Therefore, although we note that the model selection process can be improved by introducing correction techniques that lower the false positive rate, it will not be covered in this thesis but left as an indication for future research.

We also note that we can use likelihood ratio tests to perform feature selection if we prefer to consider the effect of covariates as well. The methodology will be very similar to that described in this section.

5.3 Goodness of fit test

After a specific distribution is chosen for the latency model F_0, it is important to verify whether or not the chosen parametric survival function fits the data well. In the context of the survival

analysis with long-term survivors, a formal goodness of fit test that tests the parametric form of F_0 was developed by Geerdens, Janssen and Van Keilegom (2020).

Consider the following set of hypotheses

$$\begin{aligned} &H_0\text{: The chosen parametric distribution } \textit{is} \text{ an appropriate choice for } F_0, \\ &H_1\text{: The chosen parametric distribution } \textit{is not} \text{ an appropriate choice for } F_0. \end{aligned} \tag{5.13}$$

The test statistic of the goodness of fit test is usually a measurement of the difference between the parametric fit for F_0 and the nonparametric fit based on the Kaplan-Meier estimator (assuming no covariates). For example, Maller and Zhou (1996) used the correlation as the test statistic, where a correlation close to 1 fails to reject the null. However, it is difficult to discuss the asymptotic properties that justify using the correlation as the test statistic. To fill this gap, Geerdens, Janssen and Van Keilegom (2020) used the Cramér-von Mises distance as the test statistic.

Recall $\hat{F}_n$ is the Kaplan-Meier estimator of the cumulative distribution function as defined in Theorem 1.1. The nonparametric fit for the latency model F_0 based on the Kaplan-Meier estimator is hence $\frac{\hat{F}_n(T_i)}{\hat{p}_n}$, where $\hat{p}_n = \hat{F}_n(\max(T_i))$ is the nonparametric susceptible proportion estimator under sufficient follow-up, whose properties will be discussed in Section 6.2.2. On the other hand, let $F_{0,\underset{\sim}{\theta}}$ denote the cumulative distribution function of the parametric distribution specified in H_0, where $\underset{\sim}{\theta}$ is the vector of parameters. The overall parametric mixture cure model will then have the cumulative distribution function $pF_{0,\underset{\sim}{\theta}}$, where the test uses $\hat{p}_n$, the nonparametric estimator, to estimate the susceptible proportion p. By doing so, we ensure a valid comparison between the parametric and the nonparametric fit for the latency model. The overall log-likelihood becomes

$$\log L(\underset{\sim}{\theta}) = \sum_{i=1}^{n} \left\{ C_i \log \left[\hat{p}_n f_{0,\underset{\sim}{\theta}}(T_i) \right] + (1 - C_i) \log \left[1 - \hat{p}_n F_{0,\underset{\sim}{\theta}}(T_i) \right] \right\},$$

where $f_{0,\underset{\sim}{\theta}}$ is the probability density function of the parametric distribution chosen in H_0 and T_i's are the observed survival times. The maximum likelihood estimator $\hat{\underset{\sim}{\theta}}$ is obtained by maximising the log-likelihood with respect to $\underset{\sim}{\theta}$ and the parametric fit for F_0 is $F_{0,\hat{\underset{\sim}{\theta}}}$.

With the above notation, Geerdens, Janssen and Van Keilegom (2020) defined the test statistic for the goodness of fit test to be the sum of squared differences between the parametric and

the nonparametric fit for F_0

$$\Lambda_{gof} = \sum_{i=1}^{n} \left\{ F_{0,\hat{\underset{\sim}{\theta}}}(T_i) - \frac{\hat{F}_n(T_i)}{\hat{p}_n} \right\}^2 . \tag{5.14}$$

The asymptotic distributions of the test statistic both under the null hypothesis and under local alternatives are derived in Section 3 of Geerdens, Janssen and Van Keilegom (2020). However, the authors noted that the asymptotic distributions depend on several unknown quantities and the convergence in distribution might be rather slow. They suggested using a parametric bootstrap algorithm to obtain an approximate p-value for the test. The detailed steps are given below.

1) With the given data and under H_0, compute $\hat{p}_n$, $\hat{\underset{\sim}{\theta}}$ and hence Λ_{gof}. Also compute $\hat{G}_n$, the Kaplan-Meier estimator of the censoring distribution G.

2) For the b-th bootstrap sample, generate the susceptible indicators $B_i^b \sim Bernoulli(\hat{p}_n)$ for $i = 1, 2, \cdots, n$. Generate the true survival times $T_i^{*,b}$ from $F_{0,\hat{\underset{\sim}{\theta}}}$ if $B_i^b = 1$ and set $T_i^{*,b} = \infty$ otherwise. Generate the censoring times U_i^b from $\hat{G}_n$. Set the observed survival times $T_i^b = \min(T_i^{*,b}, U_i^b)$ and the censoring indicator $C_i^b = I\{T_i^{*,b} \leq U_i^b\}$. Using the bootstrapped sample, compute the b-th bootstrap test statistic Λ_{gof}^b using (5.14).

3) Repeat **Step 2** N_b times and calculate the approximate p-value as

$$\frac{\sum_{b=1}^{N_b} I\{\Lambda_{gof}^b \geq \Lambda_{gof}\}}{N_b},$$

where a small p-value indicates evidence against the null hypothesis in (5.13).

5.4 Examples using simulated data

5.4.1 Example using simulated data with sufficient follow-up

To demonstrate implementing the methodology described in Section 5.2 in practice, we simulate the survival times of $n = 1000$ patients from a Weibull distribution with parameters $\alpha = 1.5$ and $\lambda = 1$ using the same technique that was introduced in Section 3.2. Each patient is assigned a $B_i \sim$ Bernoulli$(p = 0.75)$. The patients with $B_i = 1$ are classified as susceptibles, and the rest with $B_i = 0$ are immunes with actual survival times of ∞. The censoring distribution is assumed

to be Uniform$(0, \tau_G)$ with an additional 5% of the censoring times fixed as τ_G. In this case, we chose $r_q = 1$ because, as demonstrated in Figure 3.6, this is the minimum r_q required to avoid the problem of lack in information that is caused by insufficient follow-up.

The Kaplan-Meier estimator of the cumulative distribution F for the survival data simulated is presented in Figure 5.1. We use this simulated dataset to demonstrate the process of parametric cure model selection, and we assume a 5% significance throughout this section.

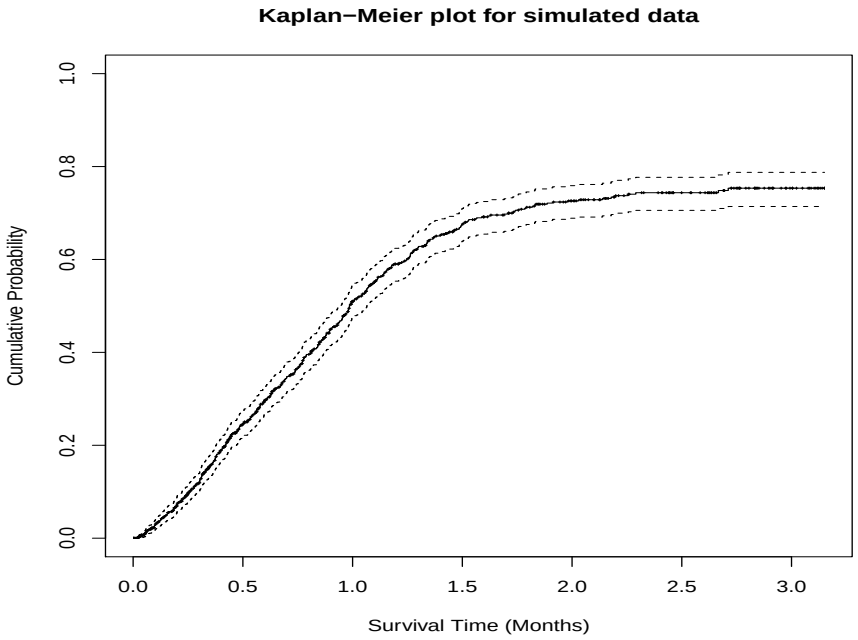


FIGURE 5.1: Plot of Kaplan-Meier estimates for survival data with $F_0 \sim$ Weibull$(1.5, 1)$ and sufficient follow-up.

Following the methodology outlined in Section 5.2, we start with **Step 1**, a test for the presence of immunes using the hypotheses in (5.2). In practice, researchers can make their decisions based on the background context of the research topic and/or the Kaplan-Meier plot. In this case, clearly, the Kaplan-Meier plot presented in Figure 5.1 levels off below 1, which suggests evidence for the presence of immunes. Researchers may also conduct a formal hypothesis test if it is of particular interest. In this section, we use Maller and Zhou's nonparametric test for the presence of immunes for demonstration. To do this, we simply compute the nonparametric estimate for the susceptible proportion, $\hat{p}_n$, and compare it with the critical value, which can be found in Maller and Zhou (1996, p.g.254 Table A.2). In this case, $\hat{p}$ is approximately 0.75, which is less than the corresponding critical value 0.9712. Hence we conclude that we should reject the null hypothesis and there is evidence for immunes in the data.

For **Step 2**, we first fit a 5-parameter generalised F mixture cure model to the data. The corresponding log-likelihood function is

$$\log L(p, \underset{\sim}{\theta}_{gf}) = \sum_{i=1}^{n} \left\{ C_i \log \left[p f_{0,\underset{\sim}{\theta}_{gf}}(T_i) \right] + (1 - C_i) \log \left[1 - p F_{0,\underset{\sim}{\theta}_{gf}}(T_i) \right] \right\}, \tag{5.15}$$

where p is the susceptible proportion. In this case, $F_{0,\underset{\sim}{\theta}_{gf}}$ and $f_{0,\underset{\sim}{\theta}_{gf}}$ are the cumulative distribution function and probability density function of the generalised F distribution with parameters $\underset{\sim}{\theta}_{gf} = \{Q, P_F, \mu_e, \sigma_e\}$ respectively. The formulation of a generalised F distribution with parameter $\{s_1, s_2, \mu, \sigma\}$ was discussed in Section 4.2.1 and its reparameterisation was discussed in Section 4.3. We use the "flexsurvcure" package in R to compute the maximum likelihood estimators for p and $\underset{\sim}{\theta}_{gf}$; we denote them as $\hat{p}$ and $\hat{\underset{\sim}{\theta}}_{gf}$ respectively. Then we are able to compute $\log L(\hat{p}, \hat{\underset{\sim}{\theta}}_{gf})$, the maximum log-likelihood under the alternative hypothesis of both (5.3) and (5.4). The results are obtained using the following line of code.

```
flexsurvcure(Surv(Time, status)~1,data=data,dist="genf", mixture=T)
```

In the above line of code, "Surv(Time, status)∼1" is the formula expression in conventional R linear modelling syntax, and the part to the left of "∼" represents the response, which must be a survival object, which is an R object defined by the "survival" package, with "Time" being the vector of survival/censoring times and "status" being the vector of the censoring indicators. The covariates are given on the right-hand side of "∼". When we write "1", no covariates are included. Examples, where covariates are included, can be found in Section 5.5.3. The "dist" argument represents the parametric distribution that should be used for F_0; "genf" corresponds to the generalised F distribution. Lastly, the "mixture=T" argument specifies that a mixture cure model should be fitted to the data. In this case, the maximised log-likelihood for the simulated dataset is

$$\log L(\hat{p}, \hat{\underset{\sim}{\theta}}_{gf}) = -891.8564.$$

Similarly, we compute the maximum log-likelihood under the null hypothesis. For example, the null hypothesis in (5.3) restricts P_F to be 0, so that a generalised gamma distribution is fitted to the susceptibles. We again use R to conduct the computations, changing dist="genf" (generalised F) to dist="gengamma", which represents the (extended) generalised gamma distribution. We fix $P_F = 0$ and maximise $\log L(\hat{p}, \underset{\sim}{\theta}_{gf}; P_F = 0)$ with respect to the parameters p, Q, μ_e and σ_e. The resulting maximum log-likelihood under the null hypothesis $(P_F = 0)$ is

-892.1277.

We then compute the log-likelihood ratio statistic using (5.1)

$$\begin{aligned}\log\left(\Lambda(P_F=0)\right) &= \log\left(\frac{L(\hat{p},\underset{\sim}{\theta}_{gf};P_F=0)}{L(\hat{p},\hat{\underset{\sim}{\theta}}_{gf})}\right)\\ &= \log L(\hat{p},\underset{\sim}{\theta}_{gf};P_F=0)-\log L(\hat{p},\hat{\underset{\sim}{\theta}}_{gf})\\ &= -892.1277+891.8564=-0.2713.\end{aligned}$$

The test statistic for the likelihood ratio test is therefore

$$-2\log(\Lambda(P_F=0))=0.5426.$$

Recall we use the 50-50 mixture of a chi-squared distribution with one degree of freedom and a point mass at 0 as the asymptotic distribution of the test statistic. In R, we simulate 10000 random variables from this distribution and use them to approximate the p-value for this test, i.e. the probability of observing a test statistic that is more extreme or as extreme as the one observed, which is calculated as the proportion of simulated values that are greater than or equal to 0.5426. In this case, the resulting p-value is

$$P(\frac{1}{2}\chi_1^2+\frac{1}{2}\delta_0\geq 0.5426)\approx 0.24,$$

where δ_0 represents a point mass at 0.

After comparing the p-value with the significance level $\alpha_s = 0.05$, we conclude that we should not reject the null hypothesis. However, we acknowledge that this does not confirm the null hypothesis is true. In this case, we cautiously choose the reduced extended generalised gamma distribution because our sample does not provide sufficient evidence to support a more complicated generalised F distribution.

At the same time, we conduct the test with hypotheses (5.4). With the log-likelihood function (5.15), we fix $Q = 0$ and maximise (5.15) with respect to p, P_F, μ_e and σ_e. By restricting Q to be 0, a Type III generalised log-logistic distribution is fitted to the susceptibles. The line of code to be used in this case is given below.

```
flexsurvcure(Surv(Time, status)~1, data=data, dist="genf",
    fixedpars=c(4), mixture=T)
```

The generalised log-logistic distribution is not an available option for the "dist" argument. Hence we manually fixed $Q = 0$ by inputting "fixedpars=c(4)", which fixes the fourth parameter Q (in the order $p, \mu_e, \sigma_e, Q, P_F$) to its initial value, which equals 0. The resulting maximised log-likelihood under the null hypothesis $(Q = 0)$ is -901.2408. Similarly, we compute the test statistic as $-2\log(\Lambda(Q = 0)) = 18.769$ and the corresponding p-value as $P(\chi^2_1 \geq 18.769) \approx 0$. The results indicate that we should not reduce F_0 from a generalised F model to a Type III generalised log-logistic model. Therefore, for **Step 2**, the overall conclusion is that F_0 should be reduced to an extended generalised gamma distribution.

Now we move on to **Step 3 a)**. For all of the four likelihood ratio tests with hypotheses (5.5)-(5.8), we follow a similar method as to the ones described above to compute their test statistics and p-values. After all, only the test with hypotheses (5.5) fails to reject the null hypothesis. Hence, we conclude that F_0 should be further reduced to a Weibull distribution, and we continue to the last step.

In **Step 4 a)**, both of the tests with hypotheses (5.10)-(5.11) suggest we should reject the null hypothesis. This means that we should end the model selection process and use a Weibull distribution as the parametric assumption for F_0.

After a Weibull distribution is chosen for the latency model F_0, we still need to conduct a goodness of fit test of the hypotheses (5.13). We will use the parametric bootstrap algorithm, which was outlined in Section 5.3, to obtain the p-value. In this case, we first compute $\hat{p}_n = 0.7435$, $\hat{\underset{\sim}{\theta}} = \{\hat{\alpha} = 1.5400, \hat{\lambda} = 1.0021\}$ and $\Lambda_{gof} = 0.0825$. Next, we create $N_b = 200$ bootstrap samples and compute the p-value for the goodness of fit test to be 0.85, indicating a Weibull distribution is an appropriate choice for F_0.

5.4.2 Example using simulated data with insufficient follow-up

We note that the model selection process explained in Section 5.2 works best with data that has sufficient follow-up. When the follow-up time is insufficient, the result can sometimes be rather misleading due to the lack of information as discussed in Section 3.2.

In this subsection, we simulate survival data with the same set-up as described in Section 5.4.1, but this time instead of choosing τ_G to be equal to the value of the largest actual survival time simulated, we let τ_G be the 90% quantile of the simulated actual survival times. In this case, the follow-up time is insufficient. A plot of the Kaplan-Meier estimator of the cumulative distribution F for such a dataset is provided below.

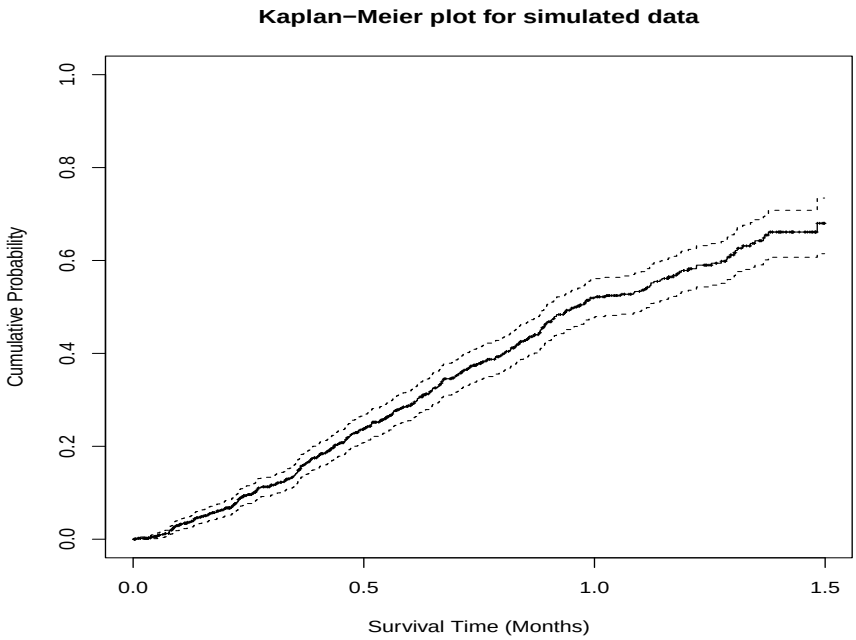


FIGURE 5.2: Plot of Kaplan-Meier estimates for survival data with $F_0 \sim$ Weibull$(1.5, 1)$ and insufficient follow-up.

For this simulated dataset, we note that Maller and Zhou's test for the presence of immunes suggests that the susceptible proportion is significantly different from 1, however, since the follow-up time is insufficient, we expect the Kaplan-Meier estimate for p to underestimate. In this case, we choose to continue the model selection process assuming the presence of immunes. The methodologies are similar to what was mentioned in Section 5.4.1. In the end, the model selection process correctly concluded that a Weibull distribution should be specified for F_0. The corresponding goodness of fit test has a p-value of 0.16, which indicates a Weibull distribution is an appropriate choice for F_0.

In addition, if we decide to fit a Weibull mixture cure model to the dataset simulated under insufficient follow-up, the maximum likelihood estimates and their corresponding standard errors (in brackets) are $\{\hat{p} = 0.8259(0.086), \hat{\alpha} = 1.4880(0.088), \hat{\lambda} = 1.0412(0.133)\}$. We note that in this case, apart from the susceptible proportion p, the maximum likelihood estimates of the parameters for the latency model are relatively accurate.

5.5 Parametric cure models applied to cancer data

The 1975-2016 database collated by the Surveillance, Epidemiology, and End Results (SEER) Program of the United States National Cancer Institute contains records of patients diagnosed

with various types of cancer from 1975 to 2016. These datasets are publicly available, and many researchers have made statistical inferences based on them. For example, we mentioned in Section 3.3.4 that Tai, Yu, Cserni, Vlastos, Royce, Kunkler and Vinh-Hung (2005) fitted a log-normal mixture cure model to the breast cancer dataset. The SEER dataset collates information from patient records collected over many years and from many sites in the United States. We treat it as observational data from which we seek to describe the characteristics of the population and use it to illustrate some of the cure model methodology and the conclusions that can be drawn from it. In this section, we restrict our analysis to male patients from the SEER database who were diagnosed with breast cancer [1].

Breast cancer is often thought of as a disease that only affects women, however, it also affects men on rare occasions. Donegan and Redlich (1996) noted that although men account for less than 1% of all breast cancer cases in the United States, it is still a serious problem. In this section, we aim to analyse the effect of covariates on the cure proportion (incident model) and the survival model (latency model) for male patients diagnosed with breast cancer. Patients with zero or unknown survival time are excluded. We also note that there are patients appearing in the datasets who have multiple entries for diagnostics at different times. For those patients, we kept only the entry with the largest survival/censoring time, since we are interested in time to death due to a specific event after maximum available follow-up. The effects of cancer staging on the cure proportion and survival time are also of interest to us, hence we have also excluded patients with unknown staging [2].

Exploratory data analysis is crucial for statistical inference as it helps decide the model to be fitted to the dataset. Though we have chosen to fit parametric mixture cure models to this data, it is still essential to conduct some analysis to check whether the follow-up time is sufficient before fitting models. In Section 5.5.1, we conduct tests for sufficient follow-up for the breast cancer dataset. We then apply the parametric cure model selection process to it. The results are presented in Section 5.5.2. After that, we conduct covariate analysis for the male breast cancer dataset, where the methodology and results are presented in Section 5.5.3.

[1] The SEER datasets classify patients who are alive or dead of other causes as censored, and they are coded with SEER cause-specific death classification = 0.

[2] The SEER datasets code breast cancer staging using Breast Adjusted AJCC 6TH Stage, with numbers between 10-24 = Stage I; 30-43 = Stage II; 50-63 = Stage III; 70-74 = Stage IV.

5.5.1 Exploratory data analysis

As mentioned in Section 3.2, insufficient follow-up greatly impacts statistical inference for cure models. Therefore a test for sufficient follow-up should be conducted as part of exploratory data analysis before fitting any cure model to the data. The results from the q_n test for sufficient follow-up (see Section 3.3) for the male breast cancer dataset are presented in Table 5.1. Recall in Section 3.3.1 we noted that the test statistic for the q_n test, N_n, follows a geometric distribution with parameter 0.25 asymptotically. Assuming the confidence level $\alpha = 0.05$, then the critical value for the test is 10. The results suggest that the male breast cancer dataset has sufficient follow-up for Stages II and III. On the other hand, the follow-up time for Stages I and IV male breast cancer datasets are insufficient. Nevertheless, as we showed in Sect 5.4.2, it may still be reasonable to fit a parametric model with appropriate caution to interpret the results.

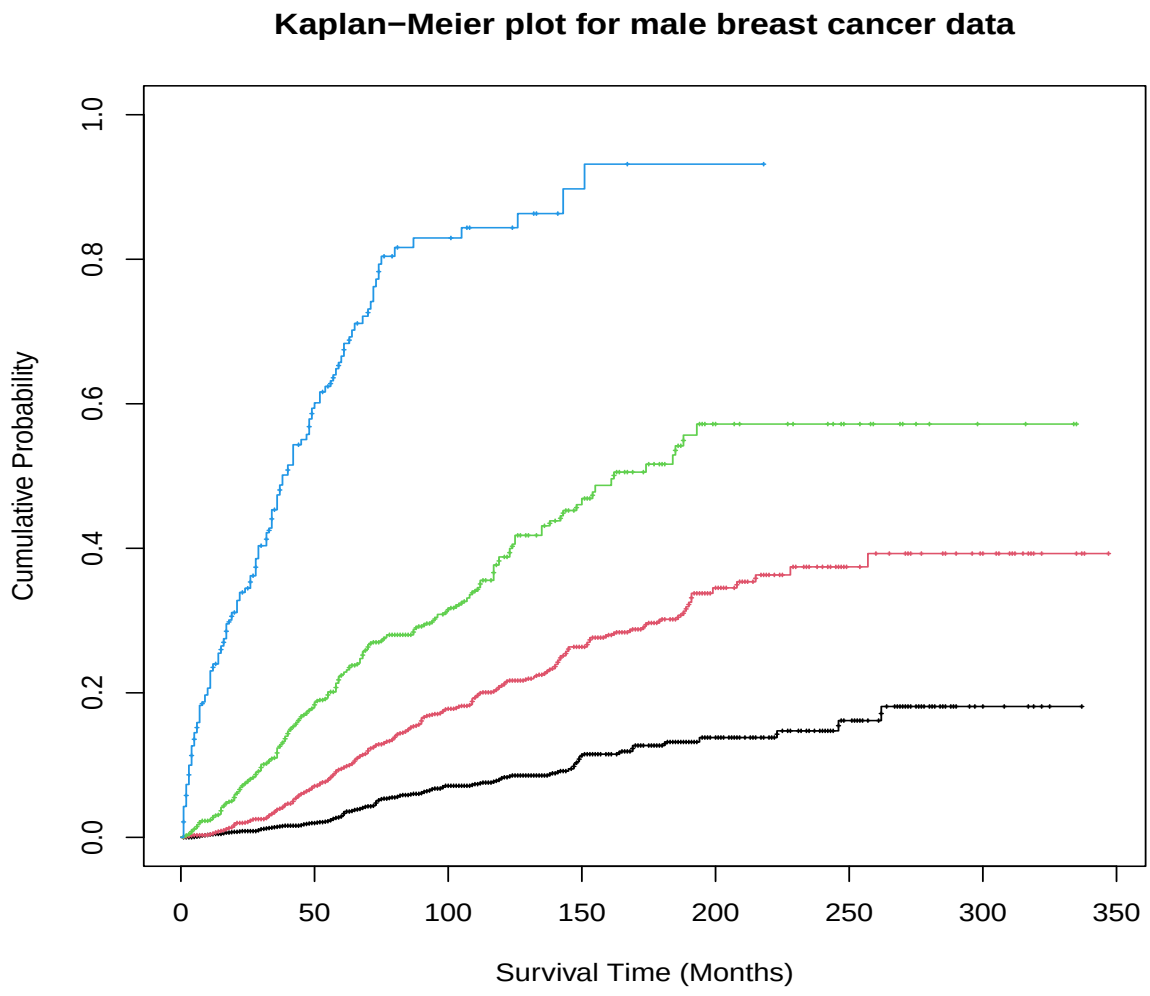


FIGURE 5.3: Kaplan-Meier plot of male breast cancer data by staging. Black = Stage I; Red = Stage II; Green = Stage III ; Blue = IV.

Staging	Sample Size (n)	Censoring (%)	N_n statistic	Follow-up
Stage I	1,087	93.01	4	Insufficient
Stage II	1,346	84.70	15	Sufficient
Stage III	577	71.92	73	Sufficient
Stage IV	234	40.17	5	Insufficient
All	3,244	82.00	21	Sufficient

TABLE 5.1: Male breast cancer data description and test statistics for the q_n test for sufficient follow-up. The critical value for the test with 95% confidence level is 10.

As observed in Figure 5.3, there appears to be a substantial cured component in the population of male patients with cancer staging I, II and III. To further support this conclusion, we conducted nonparametric tests for the presence of immune, and the results are presented in Table 5.2.

Staging	Test statistic ($\hat{p}_n$)	Critical value	Conclusion
Stage I	0.1811	0.9941	$p < 1$
Stage II	0.3927	0.9941	$p < 1$
Stage III	0.5719	0.9910	$p < 1$
Stage IV	0.9316	0.9837	$p < 1$
All	0.3739	0.9941	$p < 1$

TABLE 5.2: Male breast cancer data's test statistics & critical values (5% confidence level) for the nonparametric test for the presence of immunes.

5.5.2 Cure model selection

Before we apply the model selection process to the male breast cancer dataset, a log-rank test is conducted, which confirms survival differentials between groups of male breast cancer patients with different staging. Thus, we partition the dataset according to the variable "Stage" and conduct model selection on the resulting sub-populations. The results suggest that a log-logistic distribution should be selected as the F_0 for male patients with Stage I breast cancer, Weibull distributions should be selected for those with Stages II or III breast cancer, and an exponential distribution should be selected for those with Stage IV breast cancer. Goodness-of-fit tests suggested that the log-logistic and exponential distribution are appropriate choices for F_0

for male patients with Stage I and IV breast cancer, respectively. The tests for the Weibull distributions selected by F_0 for male patients with Stage II and III breast cancer were significant with bootstrapped p-values as 0. On the other hand, for both datasets of male patients with Stages II and III breast cancer, an extended generalised gamma distribution, which generalised the Weibull distribution, is an appropriate choice for F_0 according to the results of the goodness-of-fit tests. Therefore, the optimal model selected for F_0 should be extended generalised gamma distributions for male patients with Stages II and III cancer.

The maximum likelihood estimators of the parameters are presented in Table 5.3. For comparison, we transformed the relevant model into an equivalent generalised F model using Table 4.2, and the results are presented in Table 5.4. Figure 5.4 shows the Kaplan-Meier estimator for male patients with various cancer staging and their corresponding fitted mixture cure models. Visually, the corresponding parametric mixture cure models fit the data well for the major parts of the distributions. There are minor departures at the extreme right-hand ends of the distributions, but the discrepancies were small.

Staging	$\hat{p}$	Susceptible Survival Model
Stage I	0.281 (0.090)	Log-logistic(μ=5.322 (0.514), σ=0.610 (0.094))
Stage II	0.415 (0.057)	EGG($Q = 0.927$ (0.281), $\mu_e = 4.968$ (0.122), $\sigma_e = 0.645$ (0.116))
Stage III	0.601 (0.130)	EGG($Q = 1.144$ (0.620), $\mu_e = 4.758$ (0.127), $\sigma_e = 0.704$ (0.246))
Stage IV	0.941 (0.032)	Exponential(λ=0.020 (0.002))
All	0.614 (0.105)	Log-logistic(μ=5.306 (0.231), σ=0.917 (0.049))

TABLE 5.3: The selected mixture cure models and maximum likelihood estimates of the parameters (bootstrap standard errors in brackets). EGG stands for extended generalised gamma distribution.

Staging	Optimal F_0	$\hat{p}$	Q	P_F	μ_e	σ_e
Stage I	Log-logistic	0.281 (0.090)	0	1	5.322 (0.514)	0.431 (0.067)
Stage II	Generalised Gamma	0.415 (0.057)	0.927 (0.281)	0	4.968 (0.122)	0.645 (0.116)
Stage III	Generalised Gamma	0.601 (0.130)	1.144 (0.620)	0	4.758 (0.127)	0.704 (0.246)
Stage IV	Exponential	0.941 (0.032)	1	0	3.867 (0.112)	1
All	Log-logistic	0.614 (0.105)	0	1	5.306 (0.231)	0.648 (0.035)

TABLE 5.4: The selected models in the form of generalised F distributions (standard errors in brackets).

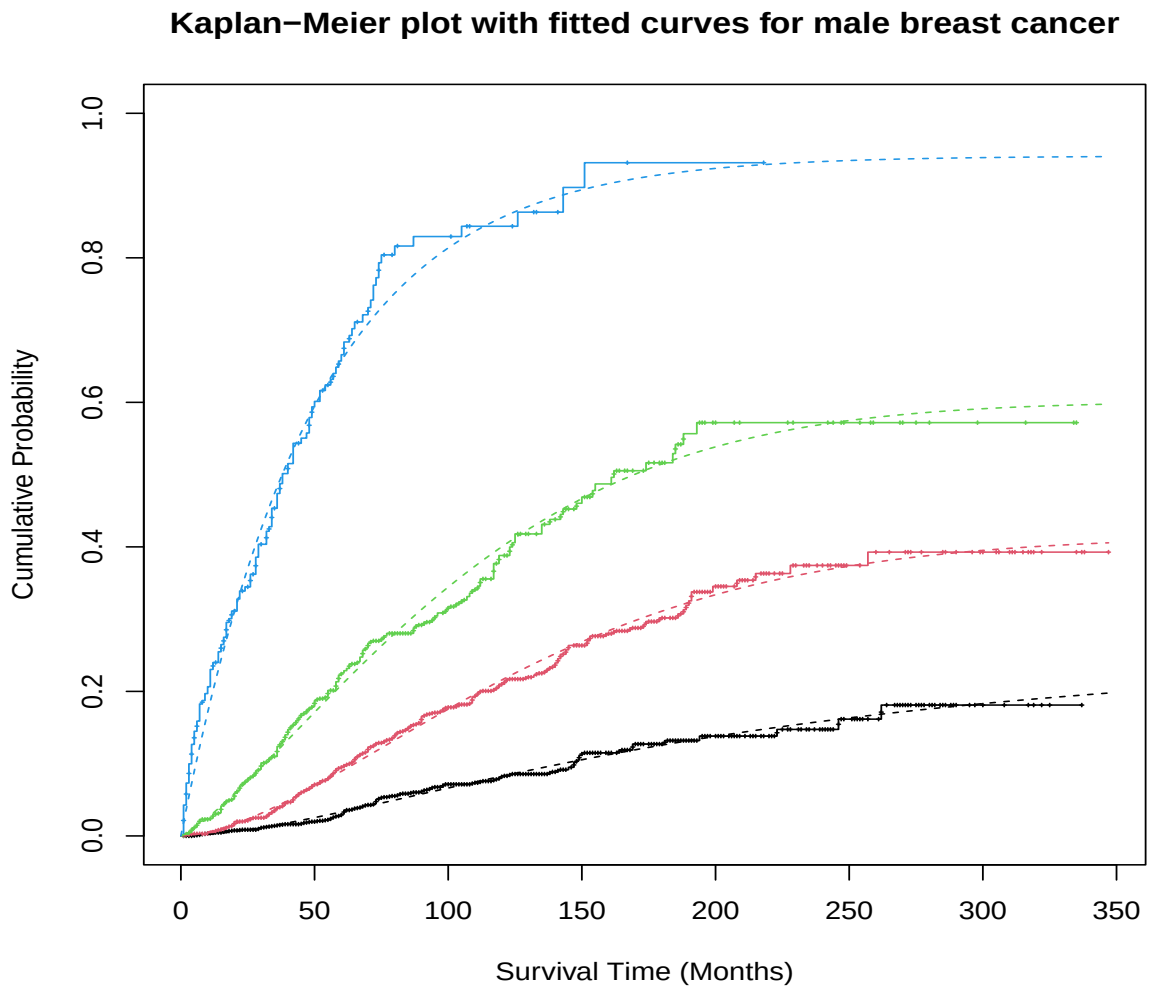


FIGURE 5.4: Kaplan-Meier plot of breast cancer data by staging. Black = Stage I; Red = Stage II; Green = Stage III; Blue = IV. Dashed line = corresponding parametric mixture cure model fit.

5.5.3 Analysis of covariate

In this section, we will focus on analysing the effect of covariates on the cure model for the male breast cancer dataset for each cancer staging. We introduce another covariate, "Genital Cancer", which is a categorical variable with two levels. Patients with "Genital Cancer"=1 have been diagnosed with both breast and genital cancer (i.e. testicular cancer for males), and for "Genital Cancer"=0, the corresponding patients have only been diagnosed with breast cancer. The covariate "Genital Cancer" is considered because breast cancer is thought to be related to genital cancers; for example, see Chandrasekaran, Scharifker, Varsegi, and Almeida (2016) for a case report of patient developing breast cancer from vaginal cancer. In this case, we are

interested in examining the significance of the effect of testicular cancer on the incident and latency models for male patients with breast cancer.

We first analyse the effect of "Genital Cancer" on the cure model selected for male patients with Stage I breast cancer. Recall that one of the advantages of mixture cure models is that the covariate, denoted as X, is allowed to have a different effect on the incidence and the latency model. In this subsection, we will first analyse the effect of "Genital Cancer" on $p(X)$, the incidence model. We follow a similar idea to that described in Section 2.2.2 and assume a logistic regression for the dependence of the cure proportion on the covariate "Genital Cancer". We write

$$p(X) = \frac{\exp(b_0 + b_1 \mathbf{1}(\text{Diagnosed with genital cancer}))}{1 + \exp(b_0 + b_1 \mathbf{1}(\text{Diagnosed with genital cancer}))},$$

where b_0 and b_1 are the coefficients, and the indicator function $\mathbf{1}(\text{Diagnosed with genital cancer})$ equals 1 if the patient is also diagnosed with testicular cancer and 0 otherwise. We can then conduct another likelihood ratio test of the hypotheses

$$H_0\text{: } b_1 = 0,$$
$$H_1\text{: } b_1 \neq 0,$$

to check whether the parameter p should be kept constant for all levels in "Genital Cancer". We note that such a test is equivalent to testing whether the cure proportion for male Stage I breast cancer patients with testicular cancer differs from that for male breast cancer patients without testicular cancer.

Recall from Table 5.3 that a log-logistic distribution should be specified for the latency model. In this case, we fit this selected cure model to data and use the R package "flexsurvcure" to compute the log-likelihood using the following line of code.

```
flexsurvcure(Surv(Time, status)~GenitalCancer, dist="llogis", mixture=T)
```

We note that the code above allows "Genital Cancer" to affect the incident model by having "GenitalCancer" on the right-hand side of $\sim$. The distribution for the latency model is specified as "llogis", which stands for a log-logistic distribution. It is then straightforward to compute the test statistic for the likelihood ratio test as

$$-2\Lambda(b_1 = 0) = -2 \times (-625.989 + 614.068) = 23.842.$$

Recall the asymptotic distribution of the test statistic is a chi-squared distribution with 1 degree of freedom. The p-value of this likelihood ratio test is hence approximately

$$P(\chi_1^2 \geq 23.842) \approx 0,$$

which indicates strong evidence against the null hypothesis. Hence the effect of "Genital Cancer" on the incident model for male patients with Stage I cancer is significant. The nonparametric estimate of the cure proportion for male patients diagnosed with both breast and genital cancer is 0.993, and the estimate for those with only breast cancer is 0.78, which is smaller. We note that, in this case, we are only interested in deaths caused by breast cancer, and by the nature of SEER cancer datasets, patients who died from other causes (including genital cancer) are considered censored in the breast cancer dataset. Thus the result, which suggests that the cure rate for patients with both breast and genital cancer is higher, might be drawn because the patients who died from genital cancer are considered cured of breast cancer. Further investigations on this issue require competing risk analysis, which will not be covered in this thesis. This is considered an indication for future research. However, we conducted a naive approach, where we manually changed the censoring indicator and the survival time of all male patients who were killed by genital cancer (there are two Stage I breast cancer patients killed by testicular cancer), so they became consistent with the information contained in the SEER genital cancer dataset. The same likelihood ratio suggests that the effect of "Genital Cancer" on the cure rate is still significant, and the estimated cure rate for these 182 patients with both cancers changed from 0.993 to 0.840, which is still higher than those 905 patients diagnosed with only breast cancer.

Following the same methodology, we analyse the effect of "Genital Cancer" on the incident model for male patients with Stages II and III breast cancer, the conclusions are the same as those with Stage I breast cancer. That is, the effect of "Genital Cancer" is significant and the cure rates for patients with both cancers are higher. For patients with Stage IV cancer, if we are only considering deaths caused by breast cancer, then the effect of "Genital Cancer" on the cure rate is significant. The result suggests that none of the ten patients with both Stage IV breast cancer and genital cancer died from breast cancer. However, we note that three of these ten patients died from genital cancer. If we change the death time and censoring indicator for these three patients in the breast cancer dataset, so the information is consistent with what was

provided in the genital cancer dataset, then the cure rate changed from 1 (no one died from breast cancer) to 0 (the patient with highest survival time died from genital cancer).

Then we analyse the effect of "Genital Cancer" on the parameter μ of the log-logistic latency model by writing

$$\mu = \beta_{\mu,0} + \beta_{\mu,1}\mathbf{1}(\text{Diagnosed with genital cancer}).$$

If $\beta_{\mu,1} = 0$, μ will be constant for all levels in "Genital Cancer". For the strictly positive parameters σ, we use log-linear models

$$\log(\sigma) = \beta_{\sigma,0} + \beta_{\sigma,1}\mathbf{1}(\text{Diagnosed with genital cancer}).$$

We can then conduct likelihood ratio tests to check whether the parameters μ and σ are constant for both levels in "Genital Cancer". In other words, we test for which of the coefficients $\beta_{\mu,1}$ and $\beta_{\sigma,1}$ are not significantly different from 0. To conduct such variable selection, we use a backward elimination technique. For details on forward and backward selection, we refer to James, Witten, Hastie and Tibshirani (2013). Namely, with b_1 already included in the model, we will first add both $\beta_{\mu,1}$ and $\beta_{\sigma,1}$ to the model and compare its maximised log-likelihood with the maximised log-likelihood of the log-logistic mixture cure model with only b_1 included. The computations can be done in R using "flexsurvcure" using the following line of code.

```
flexsurvcure(Surv(Time, status)~GenitalCancer+shape(GenitalCancer)+
    scale(GenitalCancer), dist="llogis",mixture=T)
```

We note that in R, the default parameters for a log-logistic distribution are the shape parameter $(1/\sigma)$ and the scale parameter $(\exp(\mu))$. Consider the following hypotheses

$$H_0\text{: } \beta_{\mu,1} = \beta_{\sigma,1} = 0,$$
$$H_1\text{: At least one of } \beta_{\mu,1} \text{ and } \beta_{\sigma,1} \neq 0.$$

The corresponding test statistic is

$$-2\Lambda(H_0) = -2 \times (-614.068 + 612.959) = 2.218,$$

and the p-value is

$$P(\chi^2_2 \geq 2.218) = 0.3299.$$

The likelihood ratio test shows that the effect of "Genital Cancer" on the latency model for male patients with Stage I breast cancer is not statistically significant, with a 5% significance level.

Following the same methodology, we analyse the effect of "Genital Cancer" on the latency models selected for male patients with Stages II, III and IV breast cancer. The conclusions are the same as for male patients with Stage I breast cancer. That is, for the cure models selected for all male breast cancer patients, the effects of "Genital Cancer" on the incident models are significant, but its effects on the latency models are insignificant. In other words, male patients diagnosed with both testicular and breast cancer have significantly different cure proportions than those diagnosed with only breast cancer. However, for male patients that are susceptible to breast cancer, whether they are also diagnosed with testicular cancer has no effect on the distribution of their survival times. On the other hand, we note that there is only one male patient with testicular cancer and Stage I breast cancer whose survival time is uncensored, and no male patients' survival time is uncensored for those with testicular cancer and Stage IV breast cancer. These patients have a high influence on the results, as evidenced by the standard errors of the parameters.

Chapter 6

More on Cure Proportion Estimation

6.1 Introduction

When the cure models are applied in practice, a lot of attention is given to the estimation of the cure proportion, $1 - p$. While it is relatively simple to estimate the cure proportion parametrically using the parametric cure models reviewed in Chapter 2, it can be sensitive to the parametric assumptions being made. Nonparametric methods were introduced to let the data speak for themselves. Maller and Zhou (1992) proposed a consistent nonparametric estimator for p under the mixture cure model structure. However, this estimator cannot handle covariates. Xu and Peng (2014) later proposed a nonparametric cure proportion estimator allowing for covariates. We will review these cure proportion estimators in Section 6.2.

As mentioned in Chapter 3, applying cure models to data without attention to constraints or assumptions may lead to misleading or meaningless estimates of the cure proportion. Motivated by this problem, Escobar-Bach and Van Keilegom (2019) proposed a nonparametric estimator of the cure proportion designed for use in this situation. They employed extrapolation techniques from extreme value theory, as is reasonable because long term survivors, if any, will likely be distinguished by tending to have the largest, censored survival times. This is the most important part of the data for our present considerations. We will discuss the extreme value theory in detail in Section 6.3.1. Escobar-Bach and Van Keilegom (2019) assumed that the true data generating distribution F_0 of the survival times of the susceptibles belongs to the Fréchet maximum domain of attraction and derived an estimator of the susceptible proportion that performs well for a broad range of censoring regimes and parameter values. For example, in Section 4.4, we showed that the Type II, Type III generalized log-logistic distribution and the log-logistic distribution are in the Fréchet domain of attraction. We will review their methodology and

results in Section 6.3.

Distributions with maxima attracted to the Fréchet distribution are relatively heavy-tailed and constitute a narrower class than for the Gumbel distribution. In fact, we noted in Section 4.4 that the generalized gamma distribution and its sub-models, including the Weibull and the log-normal distributions, are all in the Gumbel domain of attraction. Hence it is reasonable to ask whether the Escobar-Bach and Van Keilegom methodology can be extended to distributions in the maximum domain of attraction of the other relevant limiting extreme value distribution, the Gumbel. In Sections 6.2, 6.3 and 6.4, we modify the Escobar-Bach and Van Keilegom method to obtain a nonparametric estimator of the cure proportion for distributions in the Gumbel maximum domain of attraction, when follow-up in the sample is insufficient, and show that it performs well in simulations and in an illustrative application to some breast cancer data.

6.2 Cure proportion estimators under sufficient follow-up

6.2.1 Parametric cure proportion estimation

Using the models and methodologies outlined in Chapters 2 and 4, it is obvious that with a specific parametric assumption for F_0, a parametric estimator for p can be obtained using either likelihood or Bayesian inferences. As an example, recall in Section 2.2.2, we noted that Farewell (1982) assumed a Weibull distribution for $F_0(t)$ in (2.1); the density function, specified by a location parameter λ and a shape parameter α, is

$$f(t) = \alpha\lambda(\lambda t)^{\alpha-1}\exp\left\{-(\lambda t)^{\alpha}\right\}, \quad t, \alpha, \lambda > 0.$$

With such a parametric assumption for $F_0(t)$, it is straightforward to write the likelihood function, which is proportional to

$$\prod_{i=1}^{n}\left[p\alpha\lambda(\lambda T_i)^{\alpha-1}\exp\left\{-(\lambda T_i)^{\alpha}\right\}\right]^{C_i}\left[1-p+p\exp\left\{-(\lambda T_i)^{\alpha}\right\}\right]^{1-C_i}.$$

Consequently, an estimator for p can be obtained by maximising the above likelihood function with respect to the parameters (α, λ, p). The asymptotic distribution of the estimator for p can be derived using the general central limit theorem for maximum likelihood estimators. We refer

to Maller and Zhou (1996) p.g. 97-109 for details. The effects of covariates can also be accounted by specifying a logistic function for $p(\mathbf{X})$.

Recall in Sections 2.2 and 2.3, we used the results from Li, Taylor and Sy (2001) (Theorems 2.1, 2.2, 2.3 and 2.4) to conclude that the cure models suffer from identifiability issues when both the incident and the latency models are assumed to be nonparametric. However, as most of the mixture cure models assume a logistic regression (2.2) for the incident model, and some structure (e.g. proportional hazards model or parametric distributions) for the latency model, identifiability is less of a problem in practice under sufficient follow-up. However, when there is insufficient follow-up, cure models nearly always suffer from identifiability issues, as discussed in Section 3.2. Yu, Tiwari, Cronin, and Feuer (2007) and Peng and Taylor (2014) refer to this problem as the "near non-identifiability problem". They suggested that when there is insufficient follow-up, "the near non-identifiability may manifest itself for a particular data set in a flat likelihood surface". The consequence of such an issue is numerical problems in estimation. As evidenced in Yu et al. (2004) and in Section 3.2, in practice, the near-identifiability problem usually translates into very large variances for the cure proportion estimator.

Furthermore, Yu, Tiwari, Cronin and Feuer (2004) suggested another problem caused by insufficient follow-up: the parametric estimators of p are sensitive to the assumed parametric latency model F_0. Specifically, if the assumption for the latency model is incorrect, the parametric estimator of p would be very unstable under insufficient follow-up, but it still performs well under sufficient follow-up. On the other hand, if the true model is correctly specified, the parametric estimator of p is accurate under both sufficient and insufficient follow-up. This can be illustrated by the simulation results in Section 6.5.

6.2.2 Nonparametric cure proportion estimation under sufficient follow-up

Maller and Zhou (1992) proposed a consistent nonparametric estimator for the cure fraction $1-p$ based on the Kaplan-Meier estimator of the population distribution function F. They adopted a mixture cure model structure, defined by Definition 2.1, and assumed a general independent censoring model with right censoring. Their estimator is defined as

$$\widehat{p}_n := \hat{F}_n(t_{(n)}),$$

where $\hat{F}_n(t)$ is the Kaplan-Meier estimator of the cumulative distribution function F and $t_{(n)} := \max_{1\leq i\leq n} t_i$ is the largest observed survival time in the sample. Recall that the estimator $\hat{p}_n$ was introduced and applied for goodness-of-fit tests to parametric cure models in Section 5.3. The asymptotic properties and restrictions of the estimator are summarised in the following theorem from Maller and Zhou (1996).

Theorem 6.1 (Theorems 3.2, 3.4, 4.1 in Maller and Zhou (1996))**.** *Assume the independent and identically distributed censoring model of Section 3.3. Then as $n \to \infty$,*

$$t_{(n)} \uparrow \tau_H \quad a.s.; \quad \sup_{0\leq t\leq \tau_H} |\widehat{F}_n(t) - F(t)| \xrightarrow{\mathrm{P}} 0; \tag{6.1}$$

and also

$$\widehat{F}_n(t_{(n)}) \to F(\tau_H) I\{H(\tau_H-) < 1\} + F(\tau_H-) I\{H(\tau_H-) = 1\}.$$

Suppose in addition that F is continuous at τ_H when $\tau_H < \infty$. Then for $0 < p \leq 1$

$$\widehat{p}_n = \widehat{F}_n(t_{(n)}) \xrightarrow{p} p \text{ as } n \to \infty \text{ if and only if } \tau_{F_0} \leq \tau_G. \tag{6.2}$$

Theorem 6.1 implies that the nonparametric cure proportion estimator is problematic when the largest possible failure time is outside the follow-up period. Intuitively, this is because we have no information on what is happening to the right of τ_G with a nonparametric approach.

Another restriction of the estimator (6.2) is that it cannot account for the effect of covariates. To solve the problem, Xu and Peng (2014) used a generalized Kaplan-Meier estimator proposed by Beran (1981), which is used to estimate the survival function with covariates, instead of the Kaplan-Meier estimator. For a set of survival data $(T, C, \mathbf{X})$ observed, the generalized Kaplan-Meier estimator is given by

$$1 - \hat{F}_{GKM}(t \mid \mathbf{x}) = \mathbf{1}\left(t \leq t^*_{(n)}\right) \prod_{j=1}^{n} \left[\frac{1 - \sum_{r=1}^{n} \mathbf{1}\left(T_r \leq T_j\right) B_r(\mathbf{x})}{1 - \sum_{r=1}^{n} \mathbf{1}\left(T_r < T_j\right) B_r(\mathbf{x})}\right]^{\mathbf{1}\left(T_j\leq t; C_j=1\right)}, \tag{6.3}$$

where $\mathbf{1}(A)$ is the indicator function with $\mathbf{1}(A) = 1$ if the event A is true and $\mathbf{1}(A) = 0$ otherwise; $B_j(\mathbf{x})$ is a proper weight function for the j-th observation satisfying $B_j(\mathbf{x}) \geq 0$ and $\sum_{j=1}^{n} B_j(\mathbf{x}) = 1$. The Xu and Peng (2014) estimator is, with $\tau_{(k)}$ being the largest uncensored observation,

$$p(\mathbf{x}) = 1 - \hat{F}_{GKM}(\tau_{(k)} \mid \mathbf{x}). \tag{6.4}$$

They suggested using the Nadaraya-Watson weight function for $B(\mathbf{x})$ in (6.3)

$$B_j(\mathbf{x}) = \frac{h^{-1}K\left(\frac{\mathbf{x}_j-\mathbf{x}}{h}\right)}{\sum_{r=1}^{n} h^{-1}K\left(\frac{\mathbf{x}_r-\mathbf{x}}{h}\right)}, \quad j = 1, \ldots, n,$$

where K is a kernel density function and h is a bandwidth. Under some conditions, they showed that the estimator (6.4) is consistent.

6.3 Existing cure proportion estimator under insufficient follow-up

Recall that though the nonparametric estimator (6.2) is asymptotically consistent with sufficient follow-up, when the follow-up is insufficient, it tends to underestimate p by the factor $F_0(\tau_H)$. The idea of Escobar-Bach and Van Keilegom (2018) is to add a compensating component to $\hat{p}_n$ to offset this bias, using extrapolation techniques from extreme value theory. Before discussing their estimator, we will start this section by introducing some background concepts and theorems from extreme value theory that will be used in this chapter.

6.3.1 Background concepts from extreme value theory

Extreme value theory is a useful tool for analyzing the tail behaviour of a distribution. We refer to Resnick (1987) and de Haan & Ferreira (2007) for comprehensive treatments. In particular, the famous Fisher-Tippet theorem (de Haan & Ferreira 2007, p.6) demonstrates that there are only three possible limiting distributions for normed and centred maxima of independent and identically distributed random variables; namely, the Fréchet, the Gumbel and the reverse Weibull. Of these, only the Fréchet and Gumbel assign mass to the positive half line and are suitable candidates for our kind of survival data analysis. In the sense described, distributions attracted to an extreme value distribution are said to be in the maximum domain of attraction (hereafter, domain of attraction) of that distribution.

Motivated by de Haan (1970)'s work on regular variation and its application to the weak convergence of (univariate) sample extremes, Resnick (1987) suggested that the categorization of the domain of attraction is best understood within the framework of the theory of regularly varying functions. We define the regularly varying functions as follows.

Definition 6.1 (Regularly Varying Functions)**.** A measurable function $U : \mathbb{R}_+ \to \mathbb{R}_+$ is regularly varying at infinity with index $\rho \in \mathbb{R}$ (written as $U \in RV_\rho$), if for $x > 0$

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\rho. \tag{6.5}$$

A function satisfying (6.5) with $\rho = 0$ is slowly varying at infinity. Regular variation at zero can be defined by changing $t \to \infty$ to $t \downarrow 0$.

For example, the function x^ρ is regularly varying with $\rho \in \mathbb{R}$. The functions $\log(x)$, $\log(1 + x)$, $2 + \sin(\log\log x)$ are slowly varying. The functions $2 + \sin(x)$, $\exp(x)$ are not regularly varying.

Regular variation is the basic analytical theory underpinning extreme value theory. For details in the theory of regular variation and some extensions, we refer to Resnick (1987) Section 0.4 and de Haan and Ferreira (2007) Appendix B.

With regularly varying functions defined, we now introduce the three domains of attraction, starting with the Fréchet.

Fréchet domain of attraction

A necessary and sufficient condition for F_0 to be in the domain of attraction of the Fréchet distribution is that its right extreme is infinity (i.e. $\tau_{F_0} = \infty$) and its tail $1 - F_0$ is regularly varying at infinity with negative index $-1/\gamma_e$ (i.e. $\overline{F}_0 \in RV_{-1/\gamma_e}$) for $\gamma_e > 0$. By Definition 6.1, this means that there exists an extreme value index $\gamma_e > 0$ such that

$$\lim_{t\to\infty} \frac{\overline{F}_0(tx)}{\overline{F}_0(t)} = x^{-1/\gamma_e} \text{ for all } x > 0 \tag{6.6}$$

(Resnick 1987, Prop. 1.11, p.54; de Haan and Ferreira 2006, p.19).

Such distributions are relatively heavy-tailed, having essentially algebraic rates of tail decrease near infinity, meaning the probability of extreme values decreases more slowly, allowing for more frequent extreme events.

Weibull domain of attraction

Next, the reversed Weibull (hereafter simply referred to as the Weibull) case is also related to regular variation. A necessary and sufficient condition for F_0 to be in the domain of attraction

of the reversed Weibull distribution is that its right extreme to be finite (i.e. $\tau_{F_0} < \infty$) and $F_0^*(t) := \overline{F}_0(\tau_{F_0} - t^{-1})$ to be regularly varying at infinity with negative index $1/\gamma_e$ for $\gamma_e < 0$ as $t \to \infty$ (i.e. $F_0^*(t) \in RV_{1/\gamma_e}$). By Definition 6.1, this means that there exists an extreme value index $\gamma_e < 0$ such that

$$\lim_{t\to\infty} \frac{F_0^*(tx)}{F_0^*(t)} = \lim_{t\to\infty} \frac{\overline{F}_0(\tau_{F_0} - (tx)^{-1})}{\overline{F}_0(\tau_{F_0} - t^{-1})} = x^{1/\gamma_e}, \text{ for all } x > 0$$

(Resnick 1987, Prop. 1.13, p.59; de Haan and Ferreira 2006, p.19).

Hence it is clear that F_0 belongs to the domain of attraction of the Weibull if and only if F_0^* belongs to the domain of attraction of the Fréchet. Such distributions are relatively short-tailed, and they always have a finite endpoint, meaning there is a finite limit which values cannot exceed; hence they are not of interest to this thesis.

Gumbel domain of attraction

Lastly, a necessary and sufficient condition for F_0 to belong to the Gumbel domain of attraction is that there exists a positive function $l(t)$ such that for all $y \in \mathbb{R}$,

$$\lim_{t\uparrow\tau_{F_0}} \frac{\overline{F}_0\left(t + yl(t)\right)}{\overline{F}_0(t)} = e^{-y}$$

(Resnick 1987, p.42, de Haan and Ferreira 2006, p.19). A possible choice of $l(t)$ is

$$l(t) = \frac{1}{\overline{F}_0(t)} \int_t^{\tau_{F_0}} \overline{F}_0(x)dx, \; 0 < t < \tau_{F_0}, \tag{6.7}$$

a function depending on the unknown distribution F_0 (Resnick 1987, p.42, Eq. 1.8). In the Gumbel case, the right extreme τ_{F_0} can be finite or infinite. The Gumbel domain of attraction includes distributions with light tails, meaning the probability of observing extremely large values decreases rapidly.

We will work also within a slightly smaller class of distributions, namely, those that satisfy the von Mises condition (Proposition 4.4.1). This is a sufficient though not necessary condition to be in the domain of attraction of the Gumbel distribution, but in practice imposes very little restriction.

The above results are reformulated in a seemingly more uniform way by de Haan and Ferreira (2007).

Theorem 6.2 (Theorem 1.2.5 from de Haan and Ferreira (2007), p.21)**.** *The distribution function F_0 is in the domain of attraction of the extreme value distribution with an extreme value index $\gamma_e \in \mathbb{R}$ if and only if for any $y > 0$,*

$$\lim_{t \to \tau_{F_0}} \frac{1 - F_0(t + yl(t))}{1 - F_0(t)} = \begin{cases} (1 + \gamma_e y)^{-1/\gamma_e}, & \gamma_e \neq 0, \\ \exp(-y), & \gamma_e = 0, \end{cases} \tag{6.8}$$

where

$$l(t) = \begin{cases} \gamma_e t. & \gamma_e > 0 \text{ (Fréchet)}, \\ -\gamma_e(\tau_{F_0} - t), & \gamma_e < 0 \text{ (Weibull)}, \\ \int_t^{\tau_{F_0}} (1 - F_0(x))dx/(1 - F_0(t)), & \gamma_e = 0 \text{ (Gumbel)}. \end{cases}$$

In Chapter 4, we defined the generalised F and generalised Gamma distributions, and in Section 4.4, we noted that (6.6) is satisfied by the generalised F distribution, so the Escobar-Bach and Van Keilegom (2019) results apply. The von Mises condition (4.29) is satisfied for the generalised Gamma distribution, so the results of Section 6.4.1 apply. We recall that many commonly used lifetime distributions are obtained as sub-models of these distributions. A comprehensive list was tabulated in Table 4.3.

6.3.2 Escobar-Bach and Van Keilegom's estimator (Fréchet assumption)

Escobar-Bach and Van Keilegom (2019) assumed that the true data generating distribution F_0 belongs to the maximum domain of attraction of the Fréchet distribution; thus, F_0 satisfies (6.6) for some $\gamma_e > 0$. Under this assumption, they defined their estimator as

$$\widehat{p}_y = \widehat{p}_n + \frac{\widehat{F}_n(t_{(n)}) - \widehat{F}_n(yt_{(n)})}{\widehat{y}_\gamma - 1} \quad \text{with} \quad \widehat{y}_\gamma := \frac{\widehat{F}_n(y^2 t_{(n)}) - \widehat{F}_n(yt_{(n)})}{\widehat{F}_n(yt_{(n)}) - \widehat{F}_n(t_{(n)})},$$

where the optimal choice of $y \in (0, 1)$ is the value whose corresponding $\widehat{p}_y$ is closest to the average of a bootstrap experiment. Details are given in Section 4 of Escobar-Bach and Van Keilegom (2019). Under these assumptions (that (6.6) holds and $\tau_G < \tau_{F_0}$), they showed that $\widehat{p}_y$ is asymptotically normally distributed.

Theorem 6.3 (Theorem 3.2 from Escobar-Bach and Van Keilegom (2018))**.** *Assume insufficient follow-up, $\tau_G < \tau_{F_0}$. Assume (6.8) holds with $\gamma_e > 0$, that $F_c(\tau_G^-) < 1$, and F is differentiable. Then,*

under the assumptions of Theorem 2.2, and for any $y \in (0,1)$ *such that*

$$\frac{F\left(y^2\tau_G\right) - F\left(y\tau_G\right)}{F\left(y\tau_G\right) - F\left(\tau_G\right)} \neq 1,$$

we have

$$\sqrt{n}\left(\hat{p}_y - p_y\left(\tau_G\right)\right) \xrightarrow{d} N\left(0, \sigma^2_{y,\tau_G}\right) \quad \text{as } n \to +\infty,$$

where

$$\sigma^2_{y,\tau_e} = \sum_{i,j=0}^{2} a_i\left(\tau_G\right) a_j\left(\tau_G\right) \left\{1 - F\left(y^i\tau_G\right)\right\} \left\{1 - F\left(y^j\tau_G\right)\right\} v\left(y^{i\vee j}\tau_G\right),$$

$a_0\left(\tau_G\right) = \left(1 - b_{y,\tau_G}\right)^2, a_1\left(\tau_G\right) = -2b_{y,\tau_G}\left(1 + b_{y,\tau_G}\right), a_2\left(\tau_G\right) = b^2_{y,\tau_G}, v()$ *is defined by*

$$v(t) = \int_0^t \frac{dF(s)}{\left(1 - F(s)\right)\left(1 - F\left(s^-\right)\right)\left(1 - F_c\left(s^-\right)\right)},$$

and

$$b_{y,\tau_G} = \frac{F\left(\tau_G\right) - F\left(y\tau_G\right)}{F\left(y^2\tau_G\right) - 2F\left(y\tau_G\right) + F\left(\tau_G\right)} \overset{\tau_e \to \tau_0}{\longrightarrow} \frac{1}{1 - y^{-1/\gamma_e}}.$$

6.4 Cure proportion estimation under insufficient follow-up for distributions in the Gumbel maximum domain of attraction

We aim to obtain an analogous estimator to the Escobar-Bach and Van Keilegom estimator using extrapolation techniques from extreme value theory, but in the Gumbel case, and to derive analogous properties.

6.4.1 Methodology

Recall the necessary and sufficient condition for F_0 to belong to the maximum domain of attraction of an extreme value distribution (6.8). The $l(t)$ in the Fréchet case is simply $l(t) = t\gamma_e$, but here we have to deal with the substantially more complicated version. To estimate the function $l(t)$, we proceed as follows. Assume (6.8) and set the arbitrary y in it to be $\varepsilon / l(\tau_G - \varepsilon)$, and $t = \tau_G - \varepsilon$. This gives

$$\frac{F_0(\tau_G) - F_0(\tau_G - \varepsilon)}{1 - F_0(\tau_G - \varepsilon)} \approx 1 - \exp\left(-\frac{\varepsilon}{l(\tau_G - \varepsilon)}\right), \tag{6.9}$$

where the symbol "≈" denotes an approximation valid for $\tau_G - \varepsilon$ close to τ_G, i.e. small ε. Then repeat this procedure with an arbitrary $\tilde{\varepsilon} < \varepsilon$: replace y by $\tilde{\varepsilon}/l(\tau_G - \varepsilon)$, and again replace t by $\tau_G - \varepsilon$. Then

$$\frac{F_0(\tau_G - \varepsilon + \tilde{\varepsilon}) - F_0(\tau_G)}{1 - F_0(\tau_G - \varepsilon)} \approx 1 - \exp\big(-\frac{\tilde{\varepsilon}}{l(\tau_G - \varepsilon)}\big). \tag{6.10}$$

Dividing (6.10) by (6.9), treated as equalities, recalling that $F(t) = pF_0(t)$ (2.1), and replacing $F(t)$ by its estimator $\hat{F}_n(t)$ (Kaplan-Meier estimator), an estimator for $l(\tau_G)$ is obtained by finding the value of $\hat{l}(t_{(n)} - \varepsilon)$ such that

$$\frac{\hat{F}_n(t_{(n)} - \varepsilon + \tilde{\varepsilon}) - \hat{F}_n(t_{(n)} - \varepsilon)}{\hat{F}_n(t_{(n)}) - \hat{F}_n(t_{(n)} - \varepsilon)} = \frac{1 - \exp(-\tilde{\varepsilon}/\hat{l}(t_{(n)} - \varepsilon))}{1 - \exp(-\varepsilon/\hat{l}(t_{(n)} - \varepsilon))}. \tag{6.11}$$

After some experimentation, we found that the method is not very sensitive to the value of $\tilde{\varepsilon}$ relative to ε, and henceforth we set $\tilde{\varepsilon} = \varepsilon/2$. Then defining

$$Z_n(\varepsilon) = Z(t_{(n)} - \varepsilon) := \frac{\hat{F}_n\left(t_{(n)} - \varepsilon/2\right) - \hat{F}_n\left(t_{(n)} - \varepsilon\right)}{\hat{F}_n\left(t_{(n)}\right) - \hat{F}_n\left(t_{(n)} - \varepsilon\right)}, \tag{6.12}$$

we can with some algebra solve (6.11) for $\hat{l}(t_{(n)} - \varepsilon)$ as

$$\begin{aligned}
\frac{1 - \exp(-\tilde{\varepsilon}/\hat{l}(t_{(n)} - \varepsilon))}{1 - \exp(-\varepsilon/\hat{l}(t_{(n)} - \varepsilon))} &= Z_n(\varepsilon) \\
1 - \exp(-0.5\varepsilon/\hat{l}(t_{(n)} - \varepsilon)) &= Z_n(\varepsilon) - Z_n(\varepsilon)\left(\exp(-0.5\varepsilon/\hat{l}(t_{(n)} - \varepsilon))\right)^2 \\
1 - \exp(-0.5\varepsilon/\hat{l}(t_{(n)} - \varepsilon)) &= Z_n(\varepsilon)\left(1 + \exp(-0.5\varepsilon/\hat{l}(t_{(n)} - \varepsilon))\right)\left(1 - \exp(-0.5\varepsilon/\hat{l}(t_{(n)} - \varepsilon))\right) \\
\exp(-0.5\varepsilon/\hat{l}(t_{(n)} - \varepsilon)) &= \frac{1}{Z_n(\varepsilon)} - 1 \\
\hat{l}(t_{(n)} - \varepsilon) &= -\frac{\varepsilon}{2\log(1/Z_n(\varepsilon) - 1)}.
\end{aligned} \tag{6.13}$$

In addition, we can use the relationship (2.1) to obtain an estimator for p from (6.9), again treated as an equality, to write

$$\frac{F(\tau_G) - F(\tau_G - \varepsilon)}{p - F(\tau_G - \varepsilon)} = 1 - \exp\left(-\varepsilon/l(\tau_G - \varepsilon)\right),$$

from which we obtain

$$p = F(\tau_G - \varepsilon) + \big(F(\tau_G) - F(\tau_G - \varepsilon)\big)\big(1 - \exp\left(-\varepsilon/l(\tau_G - \varepsilon)\right)\big)^{-1}. \tag{6.14}$$

Replacing τ_G by its consistent estimator $t_{(n)}$ in (6.14) leads to

$$\begin{aligned}\hat{p}_G(n,\varepsilon) &= \hat{F}_n(t_{(n)}-\varepsilon) + \left(\hat{F}_n(t_{(n)}) - \hat{F}_n(t_{(n)}-\varepsilon)\right)\left(1-\exp\left(-\varepsilon/\hat{l}(t_{(n)}-\varepsilon)\right)\right)^{-1} \\ &:= \hat{F}_n(t_{(n)}-\varepsilon) + \hat{C}(t_{(n)},\varepsilon). \end{aligned} \tag{6.15}$$

Substituting (6.12) and (6.13) into (6.15) we can rewrite (6.15) as

$$\begin{aligned}\hat{p}_G(n,\varepsilon) &= \hat{F}_n(t_{(n)}-\varepsilon) + \left(\hat{F}_n(t_{(n)}) - \hat{F}_n(t_{(n)}-\varepsilon)\right) \Big/ \left(1-\exp\left(\frac{2\varepsilon\log(1/Z_n(\varepsilon)-1)}{\varepsilon}\right)\right) \\ &= \hat{F}_n\left(t_{(n)}-\varepsilon\right) + \left(\hat{F}_n\left(t_{(n)}\right) - \hat{F}_n\left(t_{(n)}-\varepsilon\right)\right) \Big/ \left[\left(1+\frac{1}{Z_n(\varepsilon)}-1\right)\left(1-\frac{1}{Z_n(\varepsilon)}+1\right)\right] \\ &= \hat{F}_n(t_{(n)}-\varepsilon) + \frac{\left(\hat{F}_n(t_{(n)}-\varepsilon/2) - \hat{F}_n(t_{(n)}-\varepsilon)\right)^2}{2\hat{F}_n\left(t_{(n)}-\varepsilon/2\right) - \hat{F}_n(t_{(n)}-\varepsilon) - \hat{F}_n(t_{(n)})}. \\ &:= \hat{F}_n(t_{(n)}-\varepsilon) + C(t_{(n)},\varepsilon). \end{aligned} \tag{6.16}$$

Finally, if $\hat{p}_G(n,\varepsilon) \leq \hat{p}_n$ occurs we replace $\hat{p}_G(n,\varepsilon)$ with $\hat{p}_n$. We can interpret this estimator as the addition of an estimator of the susceptible proportion based on the Kaplan-Meier estimator and a correction term.

6.4.2 Asymptotic properties of the proposed estimator

Under some reasonable assumptions, the estimator $\hat{p}_G(n,\varepsilon)$ is consistent in a certain sense for p and asymptotically normally distributed as $n \to \infty$ and $\varepsilon \downarrow 0$.

Theorem 6.4 (Consistency of $\hat{p}_G(n,\varepsilon)$). *Assume the independent and identically distributed censoring model holds with $\tau_G < \tau_{F_0}$, and assume F_0 satisfies (4.29). Then for each $\varepsilon > 0$, when $n \to \infty$,*

$$\hat{C}(t_{(n)},\varepsilon) \xrightarrow{p} C(\tau_G,\varepsilon) := \frac{\left(F\left(\tau_G-\varepsilon/2\right) - F\left(\tau_G-\varepsilon\right)\right)^2}{2F\left(\tau_G-\varepsilon/2\right) - F\left(\tau_G-\varepsilon\right) - F\left(\tau_G\right)}, \tag{6.17}$$

and as $\varepsilon \downarrow 0$,

$$C(\tau_G,\varepsilon) \to -p\frac{F_0'\left(\tau_G\right)^2}{F_0''\left(\tau_G\right)} =: C(\tau_G).$$

As $n \to \infty$, $\varepsilon \to 0$ and $\tau_G \to \tau_{F_0}$, the estimator $\hat{p}_G(n,\varepsilon)$ is consistent in the sense that it converges to $F(\tau_G) + C(\tau_G)$, where, further,

$$\lim_{\tau_G \uparrow \tau_{F_0}} F(\tau_G) + C(\tau_G) = p. \tag{6.18}$$

Proof. When $n \to \infty$, we have $\hat{F}_n(t) \to pF_0(t)$. Since F is continuous at τ_G, $C(t;\varepsilon)$ is continuous at τ_G. By (6.1) $t_{(n)} \uparrow \tau_H = \min(\tau_F, \tau_G) = \tau_G$ with insufficient follow-up. So (6.17) follows from (6.15). Next, use a Taylor expansion to write

$$F(\tau_G - \varepsilon) = F(\tau_G) - \varepsilon F'(\tau_G) + \frac{\varepsilon^2}{2} F''(\tau_G) + O(\varepsilon^3), \tag{6.19}$$

with a similar expansion for $F(\tau_G - \varepsilon/2)$. Substituting these into (6.17), and ignoring terms of order ε^3, we can obtain

$$C(\tau_G, \varepsilon) \approx -\frac{\left(\frac{\varepsilon}{2}F'(\tau_G) - \frac{3\varepsilon^2}{8}F''(\tau_G)\right)^2}{\frac{\varepsilon^2}{4}F''(\tau_G)} = -\frac{\left(F'(\tau_G) - \frac{3\varepsilon}{4}F''(\tau_G)\right)^2}{F''(\tau_G)}.$$

As $\varepsilon \to 0$, we find using (2.1) that

$$\lim_{\varepsilon \to 0} C(\tau_G, \varepsilon) = -\frac{F'(\tau_G)^2}{F''(\tau_G)} = -p\frac{F_0'(\tau_G)^2}{F_0''(\tau_G)} := C(\tau_G).$$

Finally (4.29) gives

$$\lim_{\tau_G \uparrow \tau_{F_0}} \frac{(1 - F_0(\tau_G))\, F_0''(\tau_G)}{(F_0'(\tau_G))^2} = -1 = -\lim_{\tau_G \to \tau_{F_0}} \frac{p\,(1 - F_0(\tau_G))}{C(\tau_G)}. \tag{6.20}$$

Note from this that $C(\tau_G) \sim p\,(1 - F_0(\tau_G))$ is bounded. Consequently (6.20) implies

$$\lim_{\tau_G \to \tau_{F_0}} F(\tau_G) + C(\tau_G) = p,$$

as required for (6.18). This completes the proof of Theorem 6.4. □

Next we look at the asymptotic normality of our estimator.

Theorem 6.5 (Asymptotic Normality of $\hat{p}_G(n,\varepsilon)$). *Assume the independent and identically distributed censoring model holds with $\tau_G < \tau_{F_0}$, and assume F_0 satisfies (4.29). Assume also that $\lim_{n\to\infty} n\overline{G}(\tau_H - \delta/\sqrt{n}) = \infty$ for each $\delta > 0$ and assume the integrability assumption Eq. (3.61) of Maller and Zhou (1996). Then $\hat{p}_G(n,\varepsilon)$ is asymptotically normal as $n \to \infty$ for each $\varepsilon > 0$, and the limiting distribution is normal as $\varepsilon \to 0$.*

Remark. The condition $\lim_{n\to\infty} n\overline{G}(\tau_H - \delta/\sqrt{n}) = \infty$ is satisfied in most realistic situations; for example, if G is uniform on $[0, \tau_G]$.

The proof of Theorem 6.5 can be found in Section A.2.

6.4.3 Choice of ε

The choice of ε in (6.16) is at our disposal, so we can try to optimize the performance of the estimator by our choice of ε. However, the asymptotic normality of the estimator requires $\varepsilon \downarrow 0$, so in practice, we want to choose ε small in a way that is adaptive to various practical conditions. In this subsection, we propose three methods to choose the optimal ε, and compare their performances based on some simulation results, with the simulation set-up described in Section 6.5.1.

Method 1 uses a similar idea to Escobar-Bach & Van Keilegom (2019a), namely, ε is selected in "a data-driven way by means of a quadratic errors criterion". Specifically, define $\mathcal{E} = \{0.01, 0.02, \cdots, t_{(n)}\}$ as a grid of possible values for ε and let $\mathcal{E}' = \{\varepsilon \in \mathcal{E} : 0 < \widehat{p}_G(n,\varepsilon) \leq 1\}$. Then the optimal choice of ε is taken to be

$$\varepsilon^* = \underset{\varepsilon \in \mathcal{E}'}{\operatorname{argmin}} \sum_{\varepsilon' \in \mathcal{E}'} \left\{ \widehat{p}_G(n,\varepsilon) - \widehat{p}_G(n,\varepsilon') \right\}^2.$$

This selects ε from a neighbourhood where the estimator stabilises near its limiting value. Of course, the grid $\mathcal{E}$ can be made coarser or finer as desired.

Method 2 uses the requirement $\varepsilon \downarrow 0$, and takes the smallest possible value of ε as the optimal choice. Namely,

$$\varepsilon_2^* = \inf\{\varepsilon \in \mathcal{E}' : \hat{p}_G(n,\varepsilon) > \hat{p}_n\}.$$

Method 3 uses a bootstrap approach that is similar to the method of Escobar-Bach and Van Keilegom (2019). For the $j-$th bootstrap sample, let $\hat{p}_G^{(j)}(n,\varepsilon)$ denote the corresponding estimate for the susceptible proportion, and $\hat{p}_n^{(j)}$ denote the nonparametric Kaplan-Meier estimator for the susceptible proportion. The optimal $\varepsilon^{(j)}$ for each bootstrap sample is then chosen by

$$\varepsilon^{(j)} = \inf\{\varepsilon \in \mathcal{E}' : \hat{p}_G^{(j)}(n,\varepsilon) > \hat{p}_n^{(j)}\}.$$

The overall optimal choice of ε is the one that makes $\hat{p}_G(n,\varepsilon)$ closest to the average of the bootstrap experiment, that is, with $N_b = 200$ bootstrap iterations,

$$\varepsilon_3^* = \underset{\varepsilon \in \mathcal{E}'}{\operatorname{argmin}} \left| \hat{p}_G(n,\varepsilon) - \frac{1}{N_b} \sum_{j=1}^{N_b} \hat{p}_G^{(j)}(n,\varepsilon^{(j)}) \right|.$$

For each of the $N = 1000$ datasets simulated, we computed the estimated proportion with

the optimal ε chosen by the above three methods respectively. Histograms of the computed $\hat{p}_G(n, \varepsilon)$ are presented in Section A.3, and they, from left to right, represent Methods 1, 2, and 3 respectively. Along with the graphs, we have also included the mean, the standard deviation and the mean squared error (MSE) of the 1000 proportion estimates. For an estimator with simulated values $\widehat{p}_j$, the mean squared error is defined by

$$MSE = \frac{1}{N} \sum_{j=1}^{N} \left(\widehat{p}_j - p_0 \right)^2, \tag{6.21}$$

where p_0 is the true p.

After comparison, we found that a quadratic error criterion (Method 1) works well in most cases. In addition, we note that our estimator is not very sensitive to the value of $\tilde{\varepsilon}$ in (6.10) relative to ε, hence our convenient choice $\tilde{\varepsilon} = \varepsilon/2$ in (6.12). (We tried $\tilde{\varepsilon} = \varepsilon/3$ as an alternative, getting similar results as for $\tilde{\varepsilon} = \varepsilon/2$. Other choices of $\tilde{\varepsilon} < \varepsilon$ require more complicated calculations.)

6.5 Simulation results

In this section, we use simulations to examine the performance of the proposed estimator both when we have less censoring and when we have more censoring, with the underlying survival distribution for the susceptible assumed to be the Gumbel, Weibull with shape parameter $\alpha = 0.5$, exponential, or Weibull with $\alpha = 2$.

6.5.1 Simulation set-up

We generate $N = 1000$ replications of samples of size $n = 1000$ patients' survival times. Those in the susceptible group are from a distribution that satisfies the von Mises' condition in Proposition 4.4.1. For these we used Weibull distributions with scale parameter $\lambda = 1$ and shape parameters $\alpha \in \{0.5, 1, 2\}$, or a Gumbel distribution with location parameter $\mu = 3$ and scale parameter $\sigma = 1$[1]. Note that the exponential is a Weibull distribution with shape parameter $\alpha = 1$.

As in Escobar-Bach and Van Keilegom (2018), the censoring distribution is Uniform$(0, \tau_G)$ with an additional 5% of the censoring times fixed at τ_G. Using a similar setup to theirs, the

[1]Though the support for a Gumbel distribution is $(-\infty, \infty)$, with such a parameterization there is only a probability of $1.89 \times 10^{-9} \approx 0$ of getting a negative death time, so these are positive for all practical purposes.

follow-up period is taken to be $\tau_G = (q_{0.95} - q_{0.25}) * r + q_{0.25}$, where q_a represents the $a-$th percentile of the 1000 simulated survival times. The parameter $r = 0.75$, 0.95, 1.5, 2 or 10, controls the amount of censoring. Smaller values of r represent a shorter follow-up period, while $r = 10$ represents very light censoring, approximating a sufficient follow-up scenario. For each simulated dataset we estimated p using the estimator $\widehat{p}_G(n, \varepsilon)$ in (6.16).

We note that when $Z_n(\varepsilon)$ in (6.12) approaches 0.5 from above or below, $\hat{l}\,(\tau_G - \varepsilon)$ in (6.13) approaches $+\infty$ or $-\infty$. This means that when using (6.16) to estimate p for given values of ε, occasional estimates of $\hat{p}_G(n, \varepsilon) > 1$ may be produced. We discard any such estimates. The behaviour of $Z_n(\varepsilon)$ will be discussed in Section A.1.

6.5.2 Simulation results with insufficient follow-up

For each simulated dataset, we estimated p using (6.15). The histograms of the resulting estimates when $r = 0.75$ and $r = 0.95$ are shown in Figures 6.1 and 6.2. Tables 6.1 and 6.2 show the average censoring proportion and the statistics $\widehat{\tau}_G$ and $\widehat{\tau}_{F_0}$ from the 1000 simulations for each specification of F_0 and p. Here $\widehat{\tau}_G$ is the average over the 1000 simulated lengths of the follow-up periods, and $\widehat{\tau}_{F_0}$ is the average of the largest simulated T_i^*. As can be seen, the censoring is very heavy.

		Cens. prop. (%)	$\hat{\tau}_G$	Imm. prop. (%)	$\hat{\tau}_{F_0}$
$p = 0.25$	Gumbel	91.1569	5.1464	24.9793	10.4629
	Weibull ($\alpha = 0.5$)	80.2682	6.7281	25.0049	57.5531
	Exponential	84.3689	2.3158	25.0523	7.3825
	Weibull ($\alpha = 2$)	89.2645	1.4315	25.0169	2.7277
$p = 0.5$	Gumbel	82.3148	5.1381	49.9660	10.5279
	Weibull ($\alpha = 0.5$)	60.5558	6.7380	49.9951	56.9014
	Exponential	68.8171	2.3130	49.9529	7.4507
	Weibull ($\alpha = 2$)	78.5042	1.4320	50.0039	2.7255
$p = 0.75$	Gumbel	73.4125	5.1431	75.0306	10.4793
	Weibull ($\alpha = 0.5$)	40.8623	6.7133	74.9955	57.6875
	Exponential	53.0849	2.3078	75.0490	7.4681
	Weibull ($\alpha = 2$)	67.7879	1.4311	75.0283	2.7420

TABLE 6.1: Simulated Data Summary when $r = 0.75$. "Cens. prop." = Censoring Proportion; "Imm. prop." = Immune Proportion

		Cens. prop. (%)	$\hat{\tau}_G$	$\hat{\tau}_{F_0}$
$p = 0.25$	Gumbel	89.4530	5.8260	10.4997
	Weibull ($\alpha = 0.5$)	79.4618	8.6075	57.6828
	Exponential	82.8650	2.8703	7.4331
	Weibull ($\alpha = 2$)	87.3851	1.6792	2.7102
$p = 0.5$	Gumbel	78.8596	5.8225	10.4947
	Weibull ($\alpha = 0.5$)	58.9081	8.6154	57.6489
	Exponential	65.7333	2.8695	7.5344
	Weibull ($\alpha = 2$)	74.8502	1.6788	2.7253
$p = 0.75$	Gumbel	68.2689	5.8270	10.4571
	Weibull ($\alpha = 0.5$)	38.4322	8.5591	56.8619
	Exponential	48.6610	2.8668	7.5640
	Weibull ($\alpha = 2$)	62.1516	1.6784	2.7304

TABLE 6.2: Simulated Data Summary when $r = 0.95$. "Cens. prop." = Censoring Proportion.

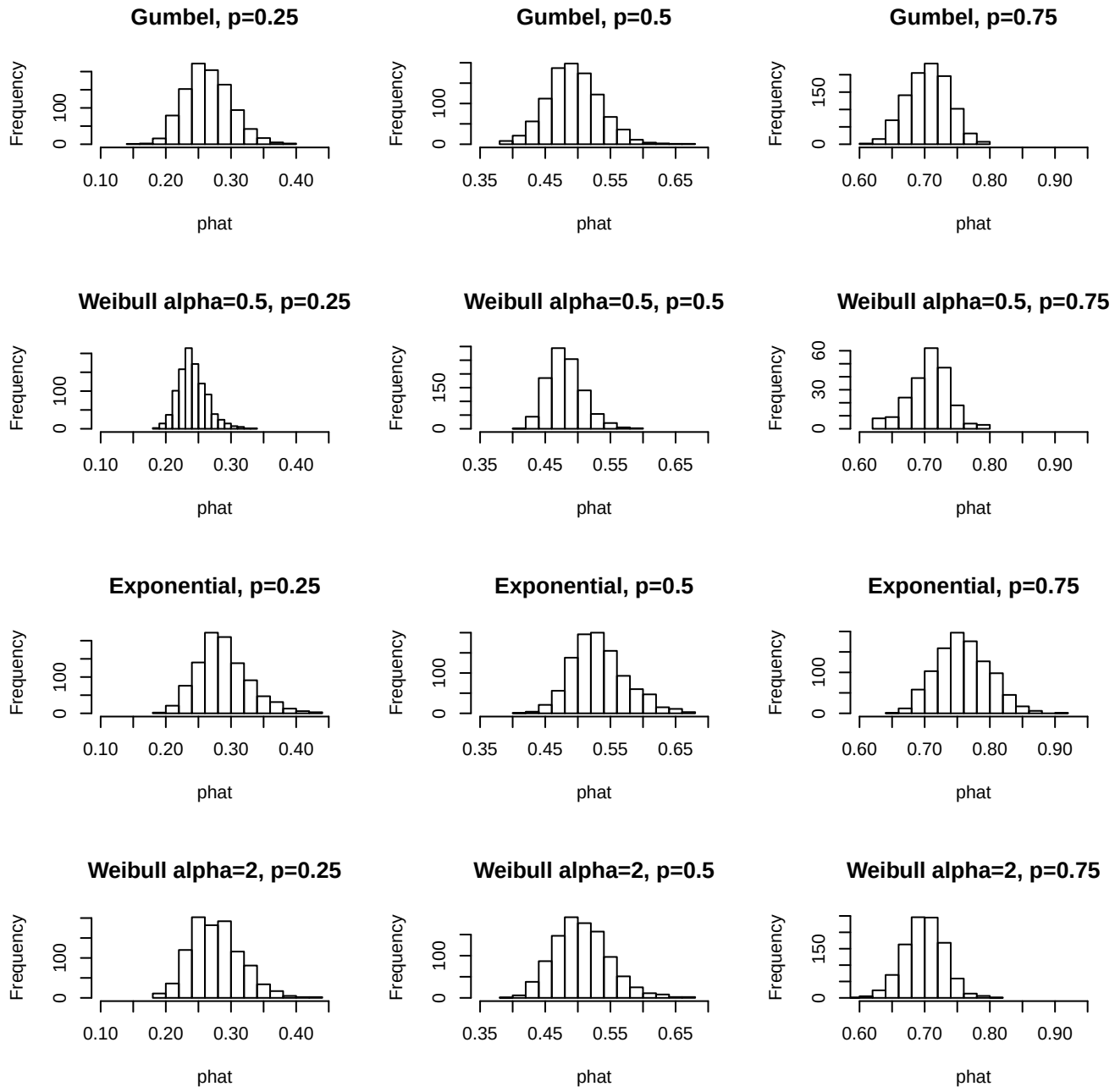


FIGURE 6.1: Histograms of estimates of p with simulated data when $r = 0.75$.

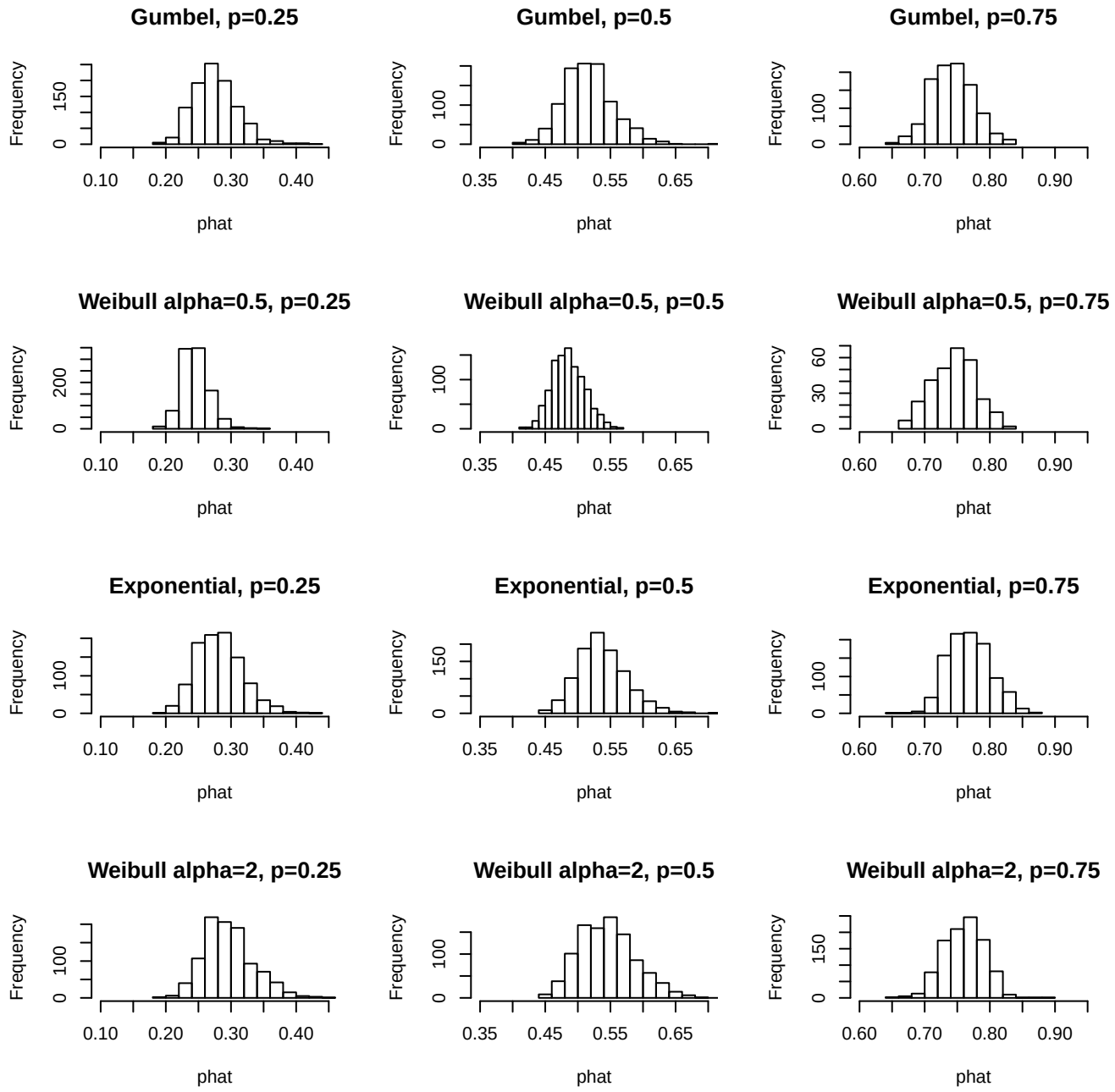


FIGURE 6.2: Histograms of estimates of p with simulated data when $r = 0.95$.

The mean, standard deviation, and mean squared error of the 1000 simulated estimates were also computed. For an estimator $\hat{p}_j$, the mean squared error is defined in the same way as Equation (6.21). We compare the performance of $\hat{p}_G(n,\varepsilon)$ with the nonparametric estimator $\hat{p}_n = \hat{F}(t_{(n)})$, and with a parametric estimator $\hat{p}_w$. The latter is obtained by fitting a Weibull mixture model to each realized set of survival times and then finding the maximum likelihood estimator $\hat{p}_w$ for p, using the method described in Section 6.2.1. Details of likelihood estimation for this model can be found in Maller and Zhou (1996) p. 97-109.

The simulation statistics from this setup are reported in Tables 6.3–6.8.

		Gumbel	Weibull ($\alpha = 0.5$)	Exponential	Weibull ($\alpha = 2$)
$\hat{p}_G(n,\varepsilon)$	MSE	0.0014	0.0006	0.0028	0.0023
	Mean	0.2629	0.2399	0.2867	0.2779
	Standard Err	0.0346	0.0216	0.0385	0.0396
$\hat{p}_n$	MSE	0.0014	0.0006	0.0011	0.0016
	Mean	0.2219	0.2300	0.2242	0.2176
	Standard Err	0.0256	0.0157	0.0199	0.0238
$\hat{p}_w$	MSE	0.0015	0.0007	0.0022	0.0076
	Mean	0.2214	0.2529	0.2554	0.2665
	Standard Err	0.0264	0.0254	0.0465	0.0858

TABLE 6.3: Simulation Summary Statistics when $r = 0.75$ and $p = 0.25$.

		Gumbel	Weibull ($\alpha = 0.5$)	Exponential	Weibull ($\alpha = 2$)
$\hat{p}_G(n, \varepsilon)$	MSE	0.0017	0.0005	0.0024	0.0038
	Mean	0.2735	0.2441	0.2826	0.2969
	Standard Err	0.0340	0.0204	0.0361	0.0400
$\hat{p}_n$	MSE	0.0008	0.0004	0.0006	0.0007
	Mean	0.2348	0.2361	0.2358	0.2350
	Standard Err	0.0236	0.0160	0.0190	0.0226
$\hat{p}_w$	MSE	0.0008	0.0005	0.0008	0.0012
	Mean	0.2320	0.2514	0.2529	0.2541
	Standard Err	0.0228	0.0220	0.0282	0.0351

TABLE 6.4: Simulation Summary Statistics when $r = 0.95$ and $p = 0.25$.

		Gumbel	Weibull ($\alpha = 0.5$)	Exponential	Weibull ($\alpha = 2$)
$\hat{p}_G(n, \varepsilon)$	MSE	0.0016	0.0011	0.0026	0.0015
	Mean	0.4957	0.4795	0.5315	0.5036
	Standard Err	0.0401	0.0265	0.0406	0.0388
$\hat{p}_n$	MSE	0.0041	0.0018	0.0031	0.0053
	Mean	0.4443	0.4623	0.4492	0.4340
	Standard Err	0.0320	0.0197	0.0234	0.0303
$\hat{p}_w$	MSE	0.0045	0.0010	0.0024	0.0072
	Mean	0.4408	0.5023	0.5051	0.5170
	Standard Err	0.0321	0.0308	0.0488	0.0830

TABLE 6.5: Simulation Summary Statistics when $r = 0.75$ and $p = 0.5$.

		Gumbel	Weibull ($\alpha = 0.5$)	Exponential	Weibull ($\alpha = 2$)
$\hat{p}_G(n, \varepsilon)$	MSE	0.0017	0.0008	0.0025	0.0037
	Mean	0.5172	0.4884	0.5342	0.5438
	Standard Err	0.0371	0.0256	0.0360	0.0416
$\hat{p}_n$	MSE	0.0017	0.0011	0.0014	0.0017
	Mean	0.4707	0.4742	0.4701	0.4687
	Standard Err	0.0287	0.0196	0.0235	0.0274
$\hat{p}_w$	MSE	0.0020	0.0007	0.0010	0.0016
	Mean	0.4640	0.5033	0.5008	0.5031
	Standard Err	0.0273	0.0271	0.0320	0.0402

TABLE 6.6: Simulation Summary Statistics when $r = 0.95$ and $p = 0.5$.

		Gumbel	Weibull ($\alpha = 0.5$)	Exponential	Weibull ($\alpha = 2$)
$\hat{p}_G(n, \varepsilon)$	MSE	0.0031	0.0020	0.0016	0.0034
	Mean	0.7054	0.7157	0.7554	0.7005
	Standard Err	0.0328	0.0292	0.0403	0.0303
$\hat{p}_n$	MSE	0.0080	0.0037	0.0063	0.0101
	Mean	0.6669	0.6927	0.6752	0.6543
	Standard Err	0.0338	0.0212	0.0256	0.0312
$\hat{p}_w$	MSE	0.0092	0.0012	0.0025	0.0057
	Mean	0.6604	0.7512	0.7545	0.7643
	Standard Err	0.0341	0.0347	0.0499	0.0741

TABLE 6.7: Simulation Summary Statistics when $r = 0.75$ and $p = 0.75$.

		Gumbel	Weibull ($\alpha = 0.5$)	Exponential	Weibull ($\alpha = 2$)
$\hat{p}_G(n, \varepsilon)$	MSE	0.0011	0.0012	0.0015	0.0010
	Mean	0.7441	0.7270	0.7695	0.7606
	Standard Err	0.0323	0.0256	0.0338	0.0305
$\hat{p}_n$	MSE	0.0028	0.0020	0.0023	0.0029
	Mean	0.7052	0.7094	0.7081	0.7042
	Standard Err	0.0284	0.0194	0.0230	0.0280
$\hat{p}_w$	MSE	0.0039	0.0008	0.0011	0.0019
	Mean	0.6938	0.7524	0.7529	0.7531
	Standard Err	0.0267	0.0282	0.0337	0.0438

TABLE 6.8: Simulation Summary Statistics when $r = 0.95$ and $p = 0.75$.

Tables 6.3–6.8 show that $\widehat{p}_G(n, \varepsilon)$ performs well in comparison with $\widehat{p}_n$, being more accurate in many cases and with comparable or even smaller mean squared error, except when $p = 0.25$ and sometimes when $p = 0.5$. In this case $\widehat{p}_G(n, \varepsilon)$ tends to be biased upwards with a somewhat higher mean squared error than $\widehat{p}_n$. The case $p = 0.25$ is the most difficult one, having high proportions both of immunes and censored data (see Table 6.1). The Kaplan-Meier estimator's largest value $\widehat{p}_n$ always underestimates p, as expected, sometimes quite markedly. The Weibull estimate $\widehat{p}_w$ works well when the underlying susceptible distribution actually is Weibull, of course, but is poor for the Gumbel distribution. Discussions on why our estimator does not perform as expected when $p = 0.25$ and $p = 0.5$, despite the asymptotic results, can be found in Appendix A.1.

6.5.3 Simulation results with sufficient follow-up

Tables 6.9-6.12 give the simulation results when $r = 10$ (very light censoring). Again each simulation has $N = 1000$ replications, each with a sample size $n = 1000$.

		Cens. prop. (%)	τ_G	Imm. prop. (%)	τ_{F_0}
$p = 0.25$	Gumbel	77.3451	35.4997	25.0661	10.4709
	Weibull ($\alpha = 0.5$)	75.5656	88.8283	24.9751	58.8154
	Exponential	75.8117	27.3400	25.0581	7.4610
	Weibull ($\alpha = 2$)	76.7216	12.4332	24.9885	2.7199
$p = 0.5$	Gumbel	54.7035	35.4863	50.0408	10.4915
	Weibull ($\alpha = 0.5$)	51.0185	88.4311	50.0764	58.4961
	Exponential	51.6767	27.3246	50.0623	7.4763
	Weibull ($\alpha = 2$)	53.4551	12.4488	49.9291	2.7392
$p = 0.75$	Gumbel	32.1615	35.4766	75.0219	10.4539
	Weibull ($\alpha = 0.5$)	26.5928	88.5623	75.0122	58.3676
	Exponential	27.6438	27.2556	74.9916	7.4562
	Weibull ($\alpha = 2$)	30.1293	12.4283	74.9657	2.7164

TABLE 6.9: Simulated Data Summary when $r = 10$.

		Gumbel	Weibull ($\alpha = 0.5$)	Exponential	Weibull ($\alpha = 2$)
$\hat{p}_G(n, \varepsilon)$	MSE	0.00020	0.00020	0.00019	0.00020
	Mean	0.24982	0.25051	0.24961	0.24984
	Standard Err	0.01404	0.01407	0.01361	0.01425
$\hat{p}_n$	MSE	0.00020	0.00020	0.00019	0.00020
	Mean	0.24982	0.25046	0.24961	0.24984
	Standard Err	0.01404	0.01406	0.01361	0.01425
$\hat{p}_w$	MSE	0.00020	0.00020	0.00019	0.00020
	Mean	0.24985	0.25053	0.24962	0.24984
	Standard Err	0.01405	0.01406	0.01361	0.01425

TABLE 6.10: Summary Statistics with sufficient follow-up and $p = 0.25$.

		Gumbel	Weibull ($\alpha = 0.5$)	Exponential	Weibull ($\alpha = 2$)
$\hat{p}_G(n, \varepsilon)$	MSE	0.00028	0.00026	0.00026	0.00028
	Mean	0.49905	0.49990	0.50004	0.50029
	Standard Err	0.01673	0.01601	0.01600	0.01674
$\hat{p}_n$	MSE	0.00028	0.00026	0.00026	0.00028
	Mean	0.49905	0.49983	0.50004	0.50029
	Standard Err	0.01673	0.01600	0.01600	0.01674
$\hat{p}_w$	MSE	0.00028	0.00025	0.00026	0.00028
	Mean	0.49912	0.49992	0.50004	0.50028
	Standard Err	0.01674	0.01596	0.01600	0.01674

TABLE 6.11: Summary Statistics with sufficient follow-up and $p = 0.5$.

		Gumbel	Weibull ($\alpha = 0.5$)	Exponential	Weibull ($\alpha = 2$)
$\hat{p}_G(n, \varepsilon)$	MSE	0.00020	0.00021	0.00020	0.00020
	Mean	0.75061	0.75052	0.75073	0.75050
	Standard Err	0.01424	0.01441	0.01396	0.01410
$\hat{p}_n$	MSE	0.00020	0.00021	0.00020	0.00020
	Mean	0.75061	0.75040	0.75073	0.75050
	Standard Err	0.01424	0.01434	0.01396	0.01410
$\hat{p}_w$	MSE	0.00020	0.00020	0.00020	0.00020
	Mean	0.75058	0.75050	0.75073	0.75050
	Standard Err	0.01422	0.01424	0.01397	0.01410

TABLE 6.12: Summary Statistics with sufficient follow-up and $p = 0.75$.

All estimators perform extremely well in this default situation – even the Weibull, when applied to data from an underlying Gumbel. Note that we are only concerned with estimating the cure proportion here, not with the goodness of fit of the distribution.

The simulation results for $r = 1.5$ and $r = 2$ are provided in Section A.5.

6.6 Breast cancer data application

It is well-known that for slowly proliferating cancers, such as early breast cancers and thyroid, the statistical cure proportions are difficult to compute due to the large number of years required for follow-up. From data sources based on the 1973-1999 Database of the Surveillance, Epidemiology, and End Results (SEER) Program of United States National Cancer Institute, Tai et al. (2005) concluded that the minimum follow-up time required for a patient with a Grade II [2] breast tumour is around 26.3 years, but records in the SEER database are less than that. Therefore, the nonparametric estimator of the cure proportion $\hat{p}_n$ is not applicable due to the problem of insufficient follow-up, but our estimator $\hat{p}_G(n, \varepsilon)$ still works in this context theoretically according to Theorem 6.4. Moreover, we investigate whether cure proportions are affected by breast cancer staging. We note that the dataset used in this section contains information on breast cancer patients with Grade II tumours for both females and males, whereas the dataset used in Section 5.5 contains information on male patients with all tumour grades.

Since the variable with information on cancer staging is only collected for patients diagnosed after 1988, we limit our analysis to the survival times of 178,505 breast cancer patients with Grade II tumours from the 1988-2016 SEER database. These are patients with information on cancer staging. Patients with zero or unknown survival time and those with unknown cancer staging are excluded. We also note that around 10% of the patients appearing in the dataset have multiple entries for diagnostics at different times. For those patients, we keep only the entry with the largest survival/censoring time for each person, since we are interested in time to death after maximum available follow-up.

Figure 6.3 shows the Kaplan-Meier estimator for each stage with fitted Weibull mixture models. Visually, follow-up appears insufficient for Stages I-III, but is probably sufficient for Stage IV. The q_n test for sufficient follow-up, which was introduced in Section 3.3, is conducted. We note that the sample sizes are all large, so we use the 95% percentile of the geometric distribution with parameter 0.25 (the asymptotic distribution of the test statistic), which is 10, as the critical value. We will reject the null hypothesis and conclude the data has sufficient follow-up if the observed nq_n is larger than 10. Otherwise, we conclude the follow-up time is insufficient. The statistics of this test for the SEER data are given in Table 6.13. The test is consistent with our impression that the follow-up period is insufficient for all cancer stages except for Stage IV.

[2] SEER program states that grade is a measurement of how closely the tumour cells resemble the parent tissue. Grade II tumour means the cells and tissue are moderately differentiated. This is an intermediate-grade tumour.

Stage	Sample Size (n)	Censoring (%)	Test statistic (N_n)
Stage I	83267	94.8	1
Stage II	55014	88.4	8
Stage III	18137	70.6	2
Stage IV	5126	38.4	86
All	161544	88.1	1

TABLE 6.13: Data description and test statistics for sufficient follow-up hypothesis test. The critical value is 10 for all tests.

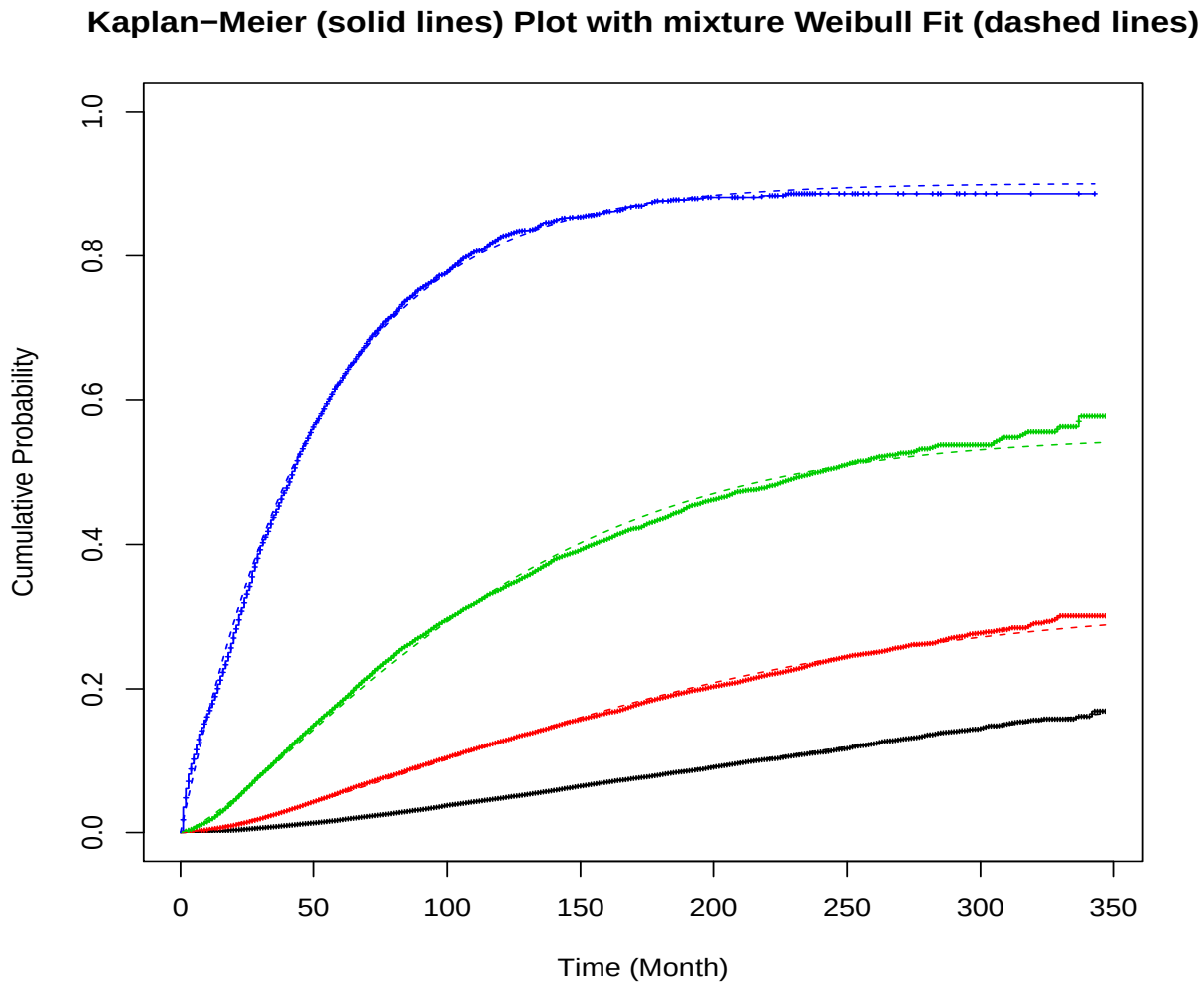


FIGURE 6.3: Breast cancer data, Kaplan-Meier plots with staging. Black, Red, Green, Blue = Stage I, II, III, IV. Dashed lines = fitted Weibull mixture models.

A Kolmogorov-Smirnov goodness-of-fit test rejects the Weibull as a model for the data, though clearly, it is a very good fit for major parts of the distributions, for each cancer stage.

There is some noticeable decrease in fit at the extreme right-hand ends of the distributions, which are of course important parts for our present analysis. (Compare with the fitted Gumbel mixture models in Section A.4.) The estimated susceptible proportions using the three estimators $\widehat{p}_G(n,\varepsilon)$, $\widehat{p}_n$, and $\widehat{p}_w$, are given in Table 6.14.

	$1-\hat{p}_G(n,\varepsilon)$	$1-\hat{p}_n$	$1-\hat{p}_w$
Stage I	0.7505 (0.0463)	0.8310	0.7363
Stage II	0.5845 (0.0399)	0.6985	0.6818
Stage III	0.3295 (0.0419)	0.4222	0.4480
Stage IV	0.1101 (0.0099)	0.1135	0.0985

TABLE 6.14: Cure proportion estimates. Bootstrap standard errors for $1-\hat{p}_G(n,\varepsilon)$ are in brackets.

The estimates using the Weibull extrapolation differ markedly (significantly in some cases) from the other two. Our proposed estimator $1-\hat{p}_G(n,\varepsilon)$ adjusts the Kaplan-Meier estimator downward, as it should, in fact quite substantially, except for Stage IV, where there is sufficient follow-up as little adjustment is needed. Generally speaking, the spread of values from the three methods gives good practical guidance as to the best estimate of the cure proportion.

Chapter 7

Conclusion and Future Work

In this thesis, we have discussed several topics in survival analysis with long-term survivals, with a particular interest in dealing with insufficient follow-up.

In Chapter 2, we reviewed some of the commonly used mixture and non-mixture cure models. We note that the parametric mixture cure models face identifiability issues under insufficient follow-up. We showed in Chapter 3 that this is because, under insufficient follow-up, the information contained in the survival data on the parameter p is significantly less than that contained by datasets with sufficient follow-up. This consequently caused the estimators of the parameters to be extremely unstable. We also reviewed some of the existing hypothesis tests and methods used to detect insufficient follow-up. However, they all have deficiencies when applied in practice.

We started Chapter 4 by introducing the generalised F distribution, which embeds most of the commonly used parametric lifetime distributions as special cases. The current literature lacks a clear outline of the relationship between the generalised F distribution and its sub-models, and we filled this gap with Tables 4.1, 4.2 and 4.3. Using these results, we outlined a model selection process for parametric mixture cure models using a generalised F distribution in Chapter 5. In Section 5.5, we demonstrated how the model selection could be applied in practice using the breast cancer dataset. Examples of covariate analysis using parametric mixture cure models are also provided in Section 5.5.3. We note that the generalised F distribution does not include the Gompertz distribution, another commonly used lifetime distribution, as its sub-model. However, we showed in Section 4.2.3 that the Gompertz distribution is closely related to the generalised F distribution. It seems possible to further generalise the generalised F distribution to contain the Gompertz distribution. This might be an area for future research.

We developed a cure proportion estimator under insufficient follow-up for distributions in

the Gumbel maximum domain of attraction in Chapter 6. We showed our estimator is asymptotically consistent and normally distributed under reasonable conditions, and it performs well in simulations. However, the effect of covariates on the cure proportion is not considered. We could consider modifying our estimator to replace the Kaplan-Meier estimators with the generalized Kaplan-Meier estimators, which allow covariates.

We note that the topic of insufficient follow-up is not well-studied in the literature, and we should aim to fill this gap in the future. We showed in Section 3.2 that for parametric mixture cure models, the problems caused by insufficient follow-up are mainly due to the flat profile log-likelihood function for the parameter p. Though the flat likelihood surface causes a lot of trouble in the frequentist approach, we wonder whether the same problem persists if we estimate the model using a Bayesian approach. We started by considering estimating the improper cumulative distribution function using a nonparametric Bayesian estimator with the Dirichlet process as the prior. However, exploratory trials showed that the mean of the resulting posterior process always tends to be a proper distribution. In the future, we suggest looking into this issue to derive a Bayesian nonparametric estimator for the survival curve and check if the identifiability issues for cure models under insufficient follow-up are still present.

Appendix A

Appendix for Chapter 6

This appendix section for Chapter 6 discusses remarks and concerns of the proposed estimator for p, proves the asymptotic normality of it, provides the plots and results for the comparison between the three methods to choose ε outlined in Section 6.4.3, gives the parametrization of the Gumbel mixture model used in simulations, and reports a variety of simulation results with differing values of r.

A.1 Remarks and concerns on the proposed estimator

As shown in Tables 6.3–6.8, we note that our estimator $\hat{p}_G(n, \varepsilon)$ might overestimate p when the follow-up time is close to being sufficient (i.e. $r = 0.95$) when F_0 follows a Weibull distribution with $\alpha = 2$. This problem is also apparent for datasets that just meet the requirement for sufficient follow-up. For illustration, we modify the simulation set-up outlined in Section 6.5.1 and introduce $r_q > 0$, with τ_G chosen to be

$$\tau_G = q_r + (q_{0.75} - q_{0.25})\,(r_q - 1)\,\mathbf{1}\,(r_q > 1)\,,$$

where q_r represents the r_q-th percentile of the simulated finite survival times and $\mathbf{1}(A)$ is a indicator function that equals 1 if the event A is true. We note that this simulation set-up is the same as what was introduced in Section 3.2. In this case, we note that $r_q = 1$ means the dataset just meets the requirement for sufficient follow-up. We simulated, with $r_q \in \{1, 3, 10\}$ (all representing sufficient follow-up), $N = 1000$ datasets with $F_0 \sim \text{Weibull}(\alpha \in \{2, 1.5, 0.5\}, \lambda = 1)$ and $p = 0.75$, and computed $\hat{p}_G(n, \varepsilon)$ for each dataset. The averages of the 1000 estimates for p are presented in Table A.1. Clearly, our proposed estimator overestimates p when $\alpha > 1$ and

the follow-up time is just sufficient, i.e. $r_q = 1$. We note when there is comfortably sufficient follow-up, i.e. $r_q = 10$, our estimator estimates p well for all α selected.

Average $\hat{p}_G(n, \varepsilon)$	$r_q = 1$	$r_q = 3$	$r_q = 10$
$\alpha = 2$	0.7730	0.7648	0.7492
$\alpha = 1.5$	0.7727	0.7565	0.7501
$\alpha = 0.5$	0.7500	0.7506	0.7502

TABLE A.1: Simulation Summary Statistics when $r = 0.75$ and $p = 0.5$.

This deficiency hinders the application of $\hat{p}_G(n, \varepsilon)$ as a test statistic for hypothesis tests. In this section, we attempt to explain why this problem arises in practice.

A.1.1 "Four-phases" behaviour of the correction term

We noted in Section 6.5.1 that $Z_n(\varepsilon)$ in (6.12) may be problematic for specific ε. In fact, after investigation, we find that $\hat{p}_G(n, \varepsilon)$ becomes unstable when forcing $\varepsilon \to 0$ because of $Z_n(\varepsilon)$. In this subsection, we use a dataset with $n = 1000$, $p = 0.75$, $r = 0.75$ that is generated by specifying an exponential distribution with $\lambda = 1$ for F_0 for illustration. The Kaplan-Meier plot of the distribution function for such a dataset is presented below.

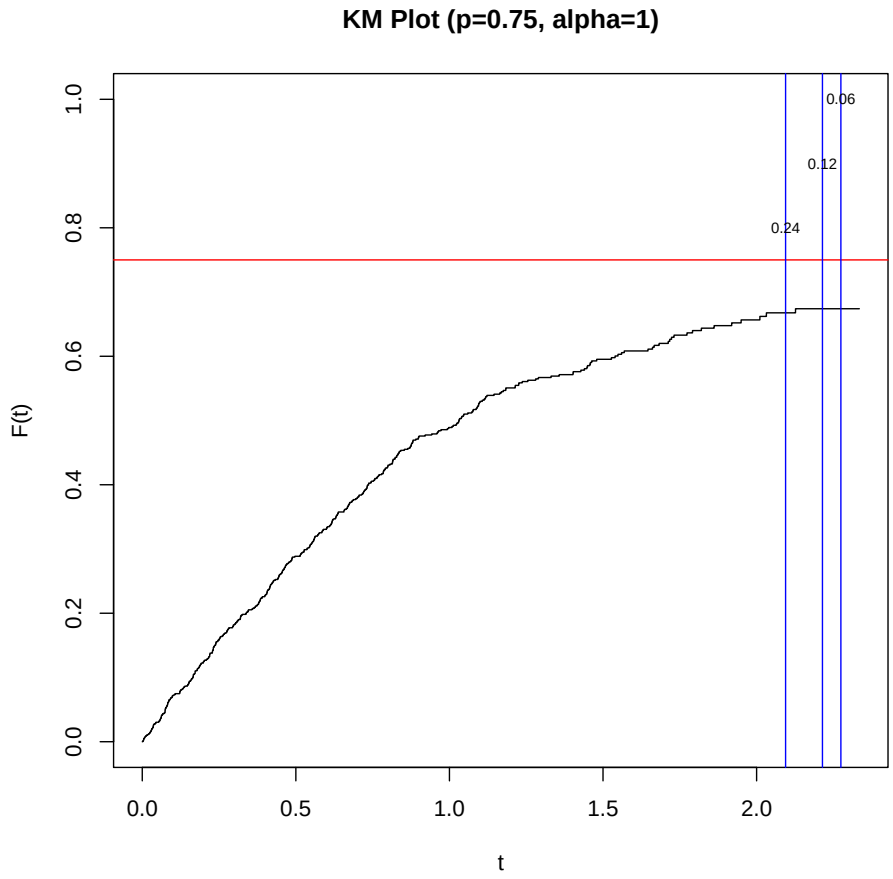


FIGURE A.1: The Kaplan-Meier plot for a dataset under sufficient follow-up simulated with $p = 0.75$ and $F_0 \sim \text{Exponential}(1)$. Here the red line represents the true proportion of susceptibles $p = 0.75$ and the blue lines represents $t = t_{(n)} - 0.24$, $t_{(n)} - 0.12$ and $t_{(n)} - 0.06$, from the left to the right respectively.

We will focus on investigating the behaviour of

$$\frac{1}{Z_n(\varepsilon)} - 1 = \frac{\hat{F}_n\left(t_{(n)}\right) - \hat{F}_n\left(t_{(n)} - \varepsilon/2\right)}{\hat{F}_n\left(t_{(n)} - \varepsilon/2\right) - \hat{F}_n\left(t_{(n)} - \varepsilon\right)}, \tag{A.1}$$

which is a key component of (6.13) and hence the overall estimator $\hat{p}_G(n, \varepsilon)$. Note that given $F_0(t) = 1 - \exp(-\lambda t)$, asymptotically,

$$\begin{aligned}
\frac{1}{Z_n(\varepsilon)} - 1 \to & \frac{1 - \exp(-\lambda t_{(n)}) - 1 + \exp(-\lambda t_{(n)} + \lambda\varepsilon/2)}{1 - \exp(-\lambda t_{(n)} + \lambda\varepsilon/2) - 1 + \exp(-\lambda t_{(n)} + \lambda\varepsilon)} \\
&= \frac{\exp(-\lambda t_{(n)} + \lambda\varepsilon/2) - \exp(-\lambda t_{(n)})}{\exp(-\lambda t_{(n)} + \lambda\varepsilon) - \exp(-\lambda t_{(n)} + \lambda\varepsilon/2)} \\
&= \frac{\exp(\lambda\varepsilon/2) - 1}{\exp(\lambda\varepsilon/2)^2 - \exp(\lambda\varepsilon/2)} \\
&= \exp(-\lambda\varepsilon/2).
\end{aligned}$$

It is clear that the discreteness of the Kaplan-Meier estimator, which can be observed in Figure A.1, affects the behaviour of (A.1). Generally, we conclude that the behaviour of Equation (A.1) with increasing ε goes through four phases. Specifically, using Figure A.1 for illustration:

1. when ε is extremely small, for example, when $\varepsilon = 0.12$ and $\varepsilon/2 = 0.06$, both the numerator and the denominator for (A.1) are 0, resulting in an NA (undefined value) for the ratio;

2. when ε is slightly larger, e.g. when $\varepsilon = 0.24$ and $\varepsilon/2 = 0.12$, the numerator for (A.1) is 0 whereas the denominator is non-zero, resulting in 0 for the ratio;

3. as ε grows larger, although not present in the above example, sometimes we also have cases where the numerator for (A.1) is non-zero, but the denominator is 0, resulting in infinity for the ratio;

4. when ε is large, both the numerator and the denominator for (A.1) are not 0, so the ratio will be close to $\exp(-\lambda\varepsilon/2)$ for large n.

The plot below illustrates the "four-phases" behaviour of (A.1), where the red line is for $\exp(-\lambda\epsilon/2)$, where $\lambda = 1$. Note that the third stage described above does not occur in this example.

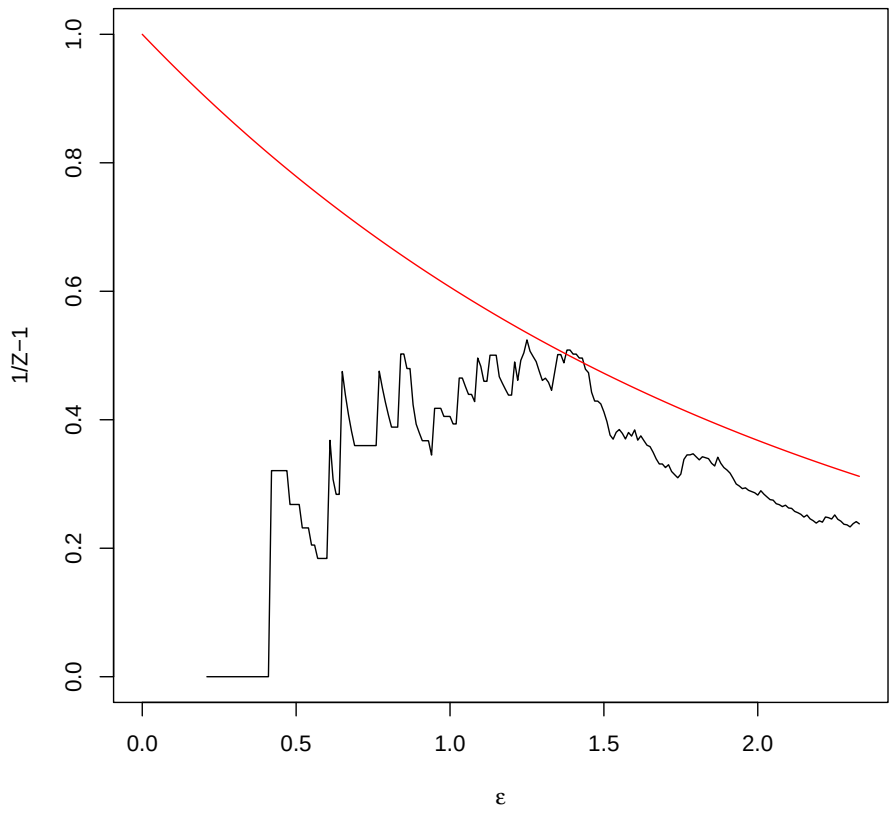


FIGURE A.2: Example plot of the "four-phases" behaviour of (A.1) using the dataset with its cumulative distribution function plotted in Figure A.1. The black line represents the calculated values of (A.1) for each ε using the Kaplan-Meier estimator $\hat{F}_n$. The red line represents the theoretical values of (A.1) if F is specified correctly.

On the other hand, we note that if the lower tail of the cumulative distribution function plot for the mixture cure model is relatively flat, for example, see Figure A.3, the behaviour of Equation (A.1) with increasing ε still goes through these four stages, but in a reversed order. For the rest of this chapter, we refer to these cases as the "flat left tail" cases.

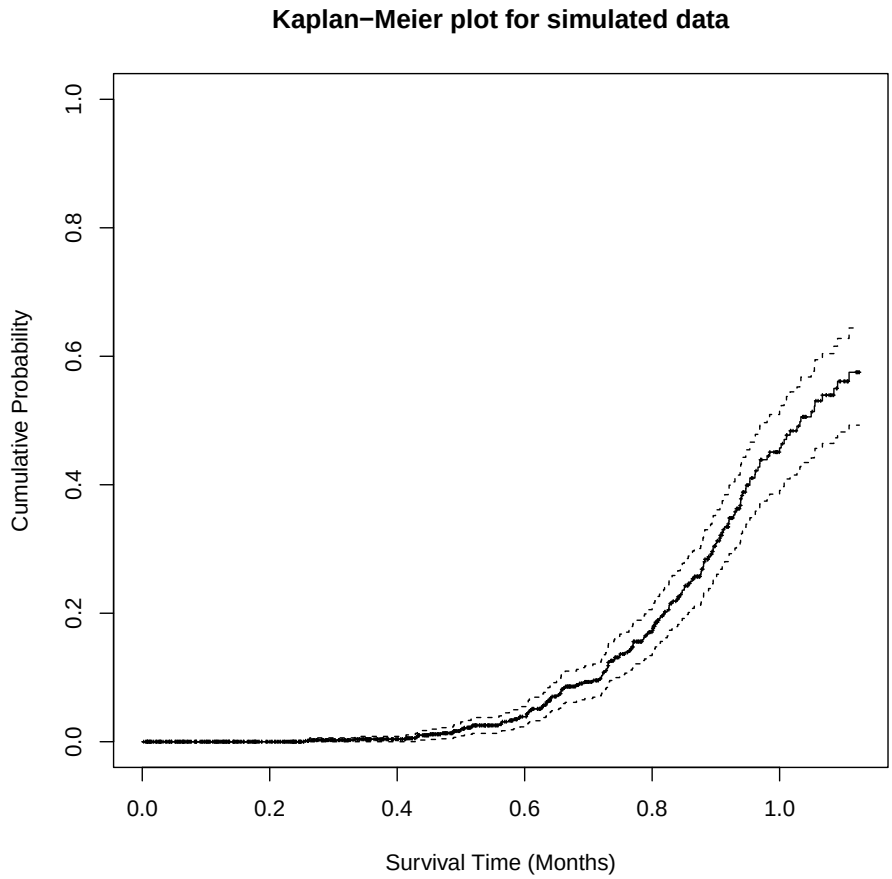


FIGURE A.3: The Kaplan-Meier plot of data under insufficient follow-up simulated from a mixture cure model with $p = 0.75$ and $F_0 \sim \text{Weibull}(\alpha = 5, \lambda = 1)$.

Overall, we conclude that, other than for the "flat left tail" cases, though theoretically we need $\varepsilon \downarrow 0$ for the estimator (6.16) to have the asymptotic properties shown in Section 6.4.2, in practice, we prefer large ε whenever possible to avoid the problems caused by the "four-phases" behaviour of the correction term. For example, for the case where F_0 is specified as an exponential distribution with $\lambda = 1$ and $p = 0.75$, when $r = 0.75$, the average optimal choice of ε for 1000 simulations is 0.7429 while the average largest susceptible survival time simulated is 1.3855; when $r = 0.95$, the average optimal ε is 2.6352 while the average largest susceptible survival time simulated is 2.9821. In addition, although it is difficult to prove, Figures 6.1 and 6.2 show that even with optimal choices of ε close to τ_G, the asymptotic distribution for $\hat{p}_n(n, \varepsilon)$ is still very likely to be normal.

Though Theorem 6.4 showed that our estimator $\hat{p}_G(n, \varepsilon)$ is always consistent when $\tau_G \to \tau_{F_0}$ and $\varepsilon \to 0$, it might also be consistent when $\tau_G \to \tau_{F_0}$ and $\varepsilon \to \tau_G$ in some special cases. In fact, we note that the optimal ε chosen in practice is sometimes close to $t_{(n)}$ instead of 0 whenever

possible due to the "four-phases" problem. More discussions on this can be found in Section A.1.2.

Furthermore, we note that specifying a parametric F_0 instead of using the Kaplan-Meier estimator $\hat{F}_n$ does not solve the problem caused by the "four-phases" behaviour. As an example, using the Weibull distribution as the specification for F_0, we perform 1000 simulations and plot the average estimates for p using $\hat{p}_G(n,\varepsilon)$, but this time we substitute the exact cumulative distribution function $F(t) = p\,(1-\exp(-\lambda^\alpha t^\alpha))$ instead of the Kaplan-Meier estimator $\hat{F}_n(t)$. For this specific simulation, we will set $\lambda = 1$, $p = 0.75$ and $r = 0.75$ for $\alpha \in (0.5, 1, 2)$ for demonstration.

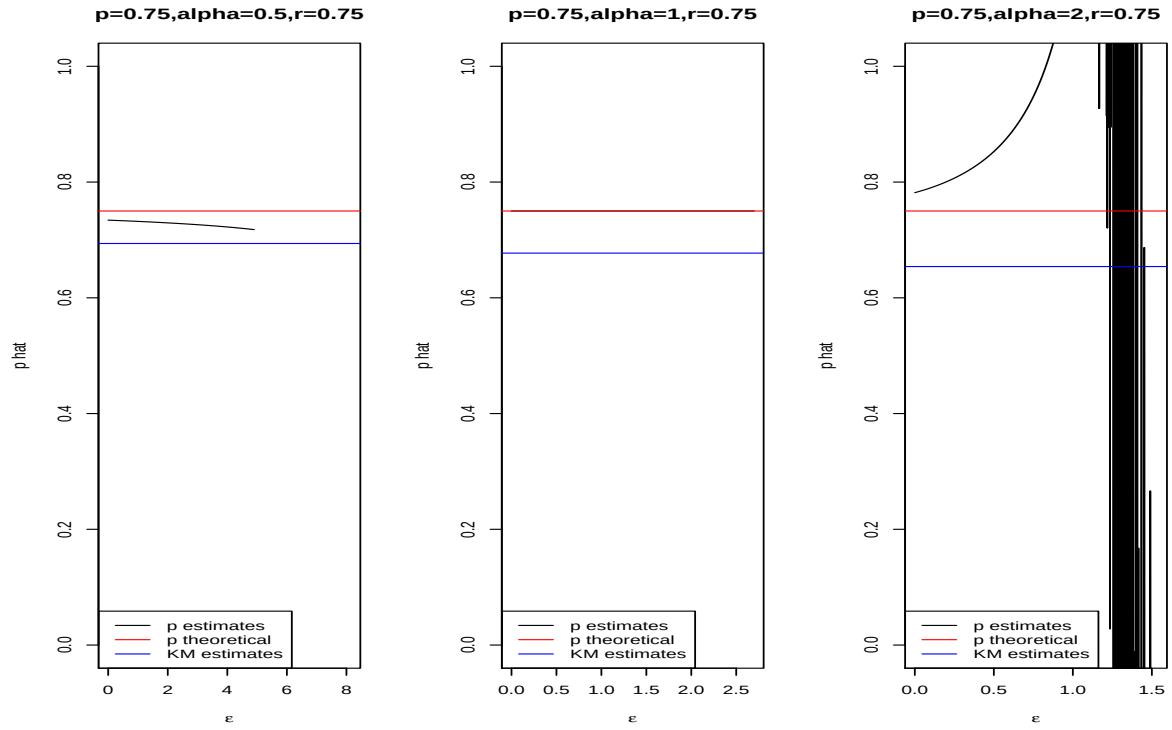


FIGURE A.4: Comparison of estimates for p with different α and hence $t_{(n)}$.

In the above plots, we note that the red horizontal line is the theoretical p, the blue horizontal line is the average $\hat{p}_n$ estimate of the susceptible proportion, and the solid black curve is the average of the estimates $\hat{p}_G(\varepsilon)$ with theoretical F, as aforementioned. We note the solid black line vanishes in the second plot where $\alpha = 1$ as it overlaps with the red line representing the true p. In the third plot, when $\varepsilon > 1$, the estimator $\hat{p}_G(\varepsilon)$ becomes extremely unstable and ranges between positive and negative infinity. The above plots suggest that when $t_{(n)}$ is small ($\alpha > 1$), the estimator does not perform up to expectation even with the theoretical $F(t)$, and it will go completely out of control as $\varepsilon \uparrow t_{(n)}$. Moreover, the plot also suggests that when $\alpha = 2$,

our estimator tends to overestimate p when the chosen ε is small; when $\alpha = 0.5$, our estimator tends to underestimate slightly.

In summary, due to the "four-phases" behaviour of the correction term of the proposed estimator, the optimal ε tends to be large in some cases, instead of being close to 0. Since ε can only be chosen between 0 and $t_{(n)}$, the problem caused by the "four-phases" behaviour is not severe when $t_{(n)}$ is large (e.g. when $\alpha < 1$ for a Weibull distribution with $p = 0.75$ and $\lambda = 1$). On the other hand, when $t_{(n)}$ is small (e.g. when $\alpha > 1$ for a Weibull mixture cure model with $p = 0.75$ and $\lambda = 1$), the choice of ε is restricted by its range, and hence forced to be small, which means the estimation for p is severely affected by the "four-phases" behaviour of the correction term.

A.1.2 Specific cases for F_0

In order to better understand the effect of having $\varepsilon \to t_{(n)}$ instead of 0 on $\hat{p}_G(n, \varepsilon)$, we replace the Kaplan-Meier estimator $\hat{F}_n(t)$ with exact cumulative distribution functions for the Gumbel, exponential, and Weibull distributions. In this section, we verify the consistency of $\hat{p}_G(\varepsilon)$ when $\varepsilon \downarrow 0$ and $\tau_G \to \tau_{F_0}$ for all of the three cases. We also show that on some occasions, $\hat{p}_G(\varepsilon) \to p$ when $\varepsilon \uparrow \tau_G$ as well.

Gumbel distribution

Firstly, assume F_0 follows a Gumbel distribution with location parameter μ and scale parameter σ, so the underlying distribution is of the form:

$$F_0(t) = \exp\left(-\exp\left(\left(\mu - t\right)/\sigma\right)\right).$$

In this case, we have

$$\begin{aligned} F_0'(t) &= \frac{\exp\left(\left(\mu - t\right)/\sigma\right)}{\sigma} \exp\left(-\exp\left(\left(\mu - t\right)/\sigma\right)\right) = \frac{\exp\left(\left(\mu - t\right)/\sigma\right)}{\sigma} F_0(t), \qquad \text{(A.2)} \\ F_0''(t) &= -\frac{\exp\left(\left(\mu - t\right)/\sigma\right)}{\sigma^2} F_0(t) + \frac{\exp\left(\left(\mu - t\right)/\sigma\right)}{\sigma} F_0'(t) \\ &= -\frac{F_0'(t)}{\sigma} + \frac{F_0'(t)}{\sigma} \exp\left(\left(\mu - t\right)/\sigma\right) \\ &= \frac{F_0'(t)}{\sigma} \left(\exp\left(\left(\mu - t\right)/\sigma\right) - 1\right). \qquad \text{(A.3)} \end{aligned}$$

Substitute (A.2) and (A.3) into the correction term $C(\tau_G, \varepsilon)$, which was defined by (6.17). When $\varepsilon \downarrow 0$, we write the correction term as

$$\begin{aligned}\lim_{\varepsilon\downarrow 0} C(\tau_G, \varepsilon) &= -p\frac{F_0'(\tau_G)^2}{F_0''(\tau_G)} \\ &= -p\frac{F_0'(\tau_G)\sigma}{\exp\left((\mu-\tau_G)/\sigma\right)-1} \\ &= -p\frac{\exp\left((\mu-\tau_G)/\sigma\right)}{\exp\left((\mu-\tau_G)/\sigma\right)-1}F_0(\tau_G).\end{aligned}$$

Hence, the estimator $\hat{p}_G(n, \varepsilon)$ defined by (6.15) becomes

$$\lim_{\varepsilon\downarrow 0}\hat{p}_G(\varepsilon) = pF_0(\tau_G) - p\frac{\exp\left((\mu-\tau_G)/\sigma\right)}{\exp\left((\mu-\tau_G)/\sigma\right)-1}F_0(\tau_G),$$

which converges to p when $\tau_G \to \tau_{F_0} = \infty$. Similarly, substitute (A.2) and (A.3) into $C(\tau_G, \varepsilon)$, when $\varepsilon \uparrow \tau_G$, to obtain

$$\begin{aligned}\lim_{\varepsilon\uparrow\tau_G} C(\tau_G, \varepsilon) &= -p\frac{\left(F_0'(\tau_G) - \frac{3\tau_G}{4}F_0''(\tau_G)\right)^2}{F_0''(\tau_G)} \\ &= -p\frac{\left(F_0'(\tau_G) - \frac{3\tau_G}{4\sigma}F_0'(\tau_G)\left(\exp\left((\mu-\tau_G)/\sigma\right)-1\right)\right)^2}{F_0'(\tau_G)\left(\exp\left((\mu-\tau_G)/\sigma\right)-1\right)/\sigma} \\ &= -pF_0'(\tau_G)\,\sigma\frac{\left(1 - \frac{3\tau_G}{4\sigma}\left(\exp\left((\mu-\tau_G)/\sigma\right)-1\right)\right)^2}{\exp\left((\mu-\tau_G)/\sigma\right)-1} \\ &= -pF_0(\tau_G)\,\frac{\exp\left((\mu-\tau_G)/\sigma\right)}{\sigma}\,\frac{\left(1 - \frac{3\tau_G}{4\sigma}\left(\exp\left((\mu-\tau_G)/\sigma\right)-1\right)\right)^2}{\exp\left((\mu-\tau_G)/\sigma\right)-1}.\end{aligned}$$

Hence, when $\varepsilon \uparrow \tau_G$ and $\tau_G \to \tau_{F_0} = \infty$, the estimator $\hat{p}_G(\varepsilon)$ becomes

$$\lim_{\varepsilon\to\tau_G}\hat{p}_G(\varepsilon) = pF_0(\tau_G - \tau_G) - pF_0(\tau_G)\,\frac{\exp\left((\mu-\tau_G)/\sigma\right)}{\sigma}\,\frac{\left(1 - \frac{3\tau_G}{4\sigma}\left(\exp\left((\mu-\tau_G)/\sigma\right)-1\right)\right)^2}{\exp\left((\mu-\tau_G)/\sigma\right)-1}, \tag{A.4}$$

where the right-hand-side is not p as $\tau_G \to \infty$. In fact, as $\tau_G \to \infty$, (A.4) approaches 0.

Therefore, we conclude if F_0 follows a Gumbel distribution, our estimator $\hat{p}_G(n, \varepsilon)$ is consistent when $n \to \infty$, $\varepsilon \downarrow 0$ and $\tau_G \to \tau_{F_0}$, however, it is not consistent when $\varepsilon \uparrow t_{(n)}$.

Exponential Distribution

Secondly, we specify an exponential distribution for F_0 with parameter λ, so that the underlying distribution is

$$F(t) = p\left(1 - \exp(-\lambda t)\right).$$

Therefore, as $n \to \infty$, (6.16) becomes

$$\begin{aligned}
\hat{p}_G(n,\varepsilon) &\to F(\tau_G - \varepsilon) + \frac{\left(F(\tau_G - \varepsilon/2) - F(\tau_G - \varepsilon)\right)^2}{2F(\tau_G - \varepsilon/2) - F(\tau_G - \varepsilon) - F(\tau_G)} \\
&= p\left(1 - \exp(-\lambda(\tau_G - \varepsilon))\right) + p\frac{\left(\exp(-\lambda\tau_G + \lambda\varepsilon) - \exp(-\lambda\tau_G + \lambda\varepsilon/2)\right)^2}{\exp(-\lambda\tau_G + \lambda\varepsilon) - 2\exp(-\lambda\tau_G + \lambda\varepsilon/2) + \exp(-\lambda\tau_G)} \\
&= p\left(1 - \exp(-\lambda(\tau_G - \varepsilon))\right) + p\exp(-\lambda\tau_G)\frac{\left(\exp(\lambda\varepsilon) - \exp(\lambda\varepsilon/2)\right)^2}{\exp(\lambda\varepsilon) - 2\exp(\lambda\varepsilon/2) + 1} \\
&= p - p\exp(-\lambda(\tau_G - \varepsilon)) + p\exp(-\lambda\tau_G)\frac{\exp(\lambda\varepsilon/2)^2\left(\exp(\lambda\varepsilon/2) - 1\right)^2}{\left(\exp(\lambda\varepsilon/2) - 1\right)^2} \\
&= p - p\exp(-\lambda(\tau_G - \varepsilon)) + p\exp(-\lambda\tau_G)\exp(\lambda\varepsilon) \\
&= p.
\end{aligned}$$

Hence, with sufficient n, $\hat{p}_G(n,\varepsilon)$ is consistent for all $\varepsilon \in [0, t_{(n)}]$, even when ε approaches τ_G.

Weibull Distribution

Lastly, assume F_0 follows a Weibull distribution with scale parameter λ and shape parameter α, so that the underlying distribution is

$$F(t) = p\left(1 - \exp\left(-(\lambda t)^\alpha\right)\right).$$

Therefore, as $n \to \infty$, $\hat{p}_G(n,\varepsilon)$ becomes

$$\begin{aligned}
\hat{p}_G(n,\varepsilon) &\to F(\tau_G - \varepsilon) + \frac{\left(F(\tau_G - \varepsilon/2) - F(\tau_G - \varepsilon)\right)^2}{2F(\tau_G - \varepsilon/2) - F(\tau_G - \varepsilon) - F(\tau_G)} \\
&= p\left(1 - \exp(-\lambda^\alpha(\tau_G - \varepsilon)^\alpha)\right) + \\
&\qquad p\frac{\left(\exp(-\lambda^\alpha(\tau_G - \varepsilon)^\alpha) - \exp(-\lambda^\alpha(\tau_G - \varepsilon/2)^\alpha)\right)^2}{\exp(-\lambda^\alpha(\tau_G - \varepsilon)^\alpha) - 2\exp(-\lambda^\alpha(\tau_G - \varepsilon/2)^\alpha) + \exp(-\lambda^\alpha\tau_G^\alpha)}.
\end{aligned} \tag{A.5}$$

When $\varepsilon \downarrow 0$,

$$\begin{aligned}\lim_{\varepsilon \downarrow 0} \hat{p}_G(\varepsilon) &= p\left(1-\exp(-\lambda^\alpha \tau_G^\alpha)\right) + p\frac{\left(\exp(-\lambda^\alpha \tau_G^\alpha) - \exp(-\lambda^\alpha \tau_G^\alpha)\right)^2}{2\left(\exp(-\lambda^\alpha \tau_G^\alpha) - \exp(-\lambda^\alpha \tau_G^\alpha)\right)} \\ &= p\left(1-\exp(-\lambda^\alpha \tau_G^\alpha)\right) \\ &:= pB(\tau_G;\lambda,\alpha), \end{aligned} \tag{A.6}$$

clearly, our estimator, as defined by (6.16), is consistent when $\tau_G \to \tau_{F_0} = \infty$. On the other hand, when $\varepsilon \uparrow \tau_G$, the limit for (A.5) becomes

$$\begin{aligned}\lim_{\varepsilon \to \tau_G} \hat{p}_G(n,\varepsilon) &= p\frac{\left(1-\exp(-2^{-\alpha}\lambda^\alpha \tau_G^\alpha)\right)^2}{1-2\exp(-2^{-\alpha}\lambda^\alpha \tau_G^\alpha) + \exp(-\lambda^\alpha \tau_G^\alpha)} \\ &:= pB_1(\tau_G;\lambda,\alpha). \end{aligned} \tag{A.7}$$

Again, when $\tau_G \to \tau_{F_0} = \infty$, the estimator $\hat{p}_G(n,\varepsilon)$ is consistent.

Overall, we conclude that in addition to Theorem 6.4, our estimator $\hat{p}_G(n,\varepsilon)$ is also consistent as $n \to \infty$, $\tau_G \to \tau_{F_0}$ and $\varepsilon \to \tau_G$ in some special cases, for example, when F_0 is specified as a Weibull distribution. However, $\hat{p}_G(n,\varepsilon)$ is not always consistent as $n \to \infty$, $\tau_G \to \tau_{F_0}$ and $\varepsilon \to \tau_G$, for example, when F_0 is specified as a Gumbel distribution.

A.1.3 The effect of $t_{(n)}$ on the estimator

To better understand the effect of having $\varepsilon \uparrow t_{(n)}$ instead of 0 on the estimator $\hat{p}_G(n,\varepsilon)$, we assign a Weibull distribution for F_0 as an example, and first plot the scaling function $B_1(t_{(n)};\lambda,\alpha)$ in (A.7). Recall that we are expecting $B_1(t_{(n)};\lambda,\alpha)$ to be as close to 1 as possible.

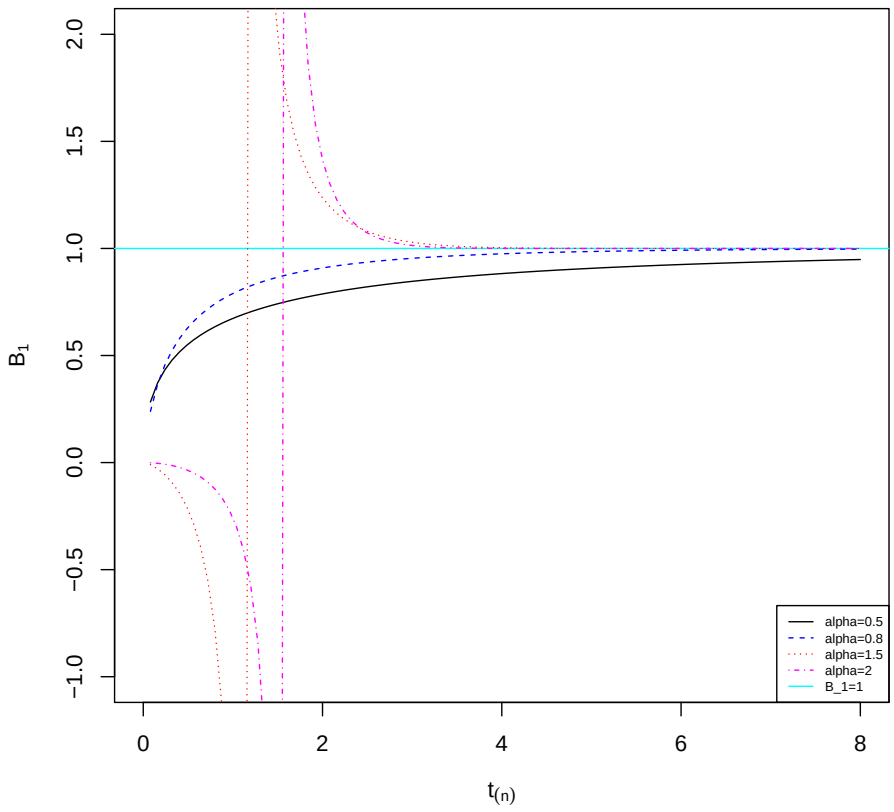


FIGURE A.5: Plot of $B_1(t_{(n)}; \lambda = 1, \alpha)$

Note that when $\alpha > 1$, under the simulation set-up as outlined in Section 6.5.1, $t_{(n)}$ is generally quite small. Using $\alpha = 2, \lambda = 1$ as an example, the simulation results suggest that $t_{(n)}$ rarely exceeds 2.5. On the other hand, when $\alpha < 1$, $t_{(n)}$ is usually quite large, for example, when $\alpha = 0.5, \lambda = 1$, the simulated $t_{(n)}$'s are usually larger than 5.

From Figure A.5, when $\alpha < 1$, we expect $t_{(n)}$ to be large, and hence $B_1(t_{(n)}; \lambda, \alpha)$ will be slightly smaller but close to 1. Therefore, our estimator will work (while it may slightly underestimate p as shown in Figure A.4) for Weibull distributions when $\alpha < 1$ if the optimal ε is close to $t_{(n)}$. On the other hand, we observe that the scaling function $B_1(t_{(n)}; \lambda, \alpha)$ evaluated at $t_{(n)} \leq 2.5$ is always greater than 1 for $\alpha > 1$ (we exclude estimates for p that are less than 0). In other words, our proposed estimator tends to overestimate p if the underlying distribution is a Weibull distribution with $\alpha > 1$, unless the level of censoring is too high for our estimator to work (i.e. when the distance between τ_G and τ_{F_0} is too large). This problem is apparent not only

for datasets under insufficient follow-up, but also those just meeting the requirement for sufficient follow-up, as shown in Table A.1. We note when there are excess amount of follow-up, i.e. $r_q = 10$, our estimator estimates p well for all α selected. This is because when r_q is large, $t_{(n)}$ will also be large, making $B_1(t_{(n)}; \lambda, \alpha)$ close to 1.

A.2 Asymptotic Normality

A main tool in the proof of Theorem 6.5 is a result of Gill (1980) giving the asymptotic distribution of the Kaplan-Meier estimator, considered as a stochastic process, in the form of a functional limit theorem. See Theorem 3.14 of Maller and Zhou (1996) (also used by Escobar-Bach and Van Keilegom (2018) in their proof). Assuming F is continuous at τ_H in case $\tau_H < \infty$, and assuming a further integrability assumption (Eq. (3.61) of Maller and Zhou (1996)), it states that the process

$$X_n(t) := \sqrt{n}\frac{\overline{F}(t)}{\overline{F}(t \wedge t_{(n)})}\Big(\hat{F}_n(t) - F(t)\Big) \tag{A.8}$$

converges in the function space $D[0, \tau_H]$ to the process $(\overline{F}(t)Z(t))_{t\geq 0}$ as $n \to \infty$. Here $(Z(t))$ is a mean zero Gaussian process with independent increments and covariance function $\mathrm{Cov}(Z(t), Z(s)) = v(t \wedge s)$ for $s, t > 0$, where

$$v(t) := \int_0^t \frac{F'(s)\mathrm{d}s}{(\overline{F}(s)^2\overline{G}(s)}. \tag{A.9}$$

The process X_n satisfies

$$X_n(t) = \begin{cases} \sqrt{n}\big(\hat{F}_n(t) - F(t)\big), \ t \leq t_{(n)}, \\ \\ \sqrt{n}\big(\hat{F}_n(t_{(n)}) - F(\tau_H)\big) + o_P(1), \ t \geq t_{(n)}. \end{cases} \tag{A.10}$$

Now to prove Theorem 6.5. For the first part, hold $\varepsilon > 0$ fixed and rewrite (6.16) as

$$\hat{p}_G(n, \varepsilon) = \frac{\hat{F}_n^2(t_{(n)} - \varepsilon/2) - \hat{F}_n(t_{(n)})\hat{F}_n(t_{(n)} - \varepsilon)}{2\hat{F}_n(t_{(n)} - \varepsilon/2) - \hat{F}_n(t_{(n)} - \varepsilon) - \hat{F}_n(t_{(n)})} =: \frac{A_n(\varepsilon)}{B_n(\varepsilon)}, \tag{A.11}$$

and define

$$A(\varepsilon) = F^2(\tau_H - \varepsilon/2) - F(\tau_H)F(\tau_H - \varepsilon) \tag{A.12}$$

and

$$B(\varepsilon) = 2F(\tau_H - \varepsilon/2) - F(\tau_H) - F(\tau_H - \varepsilon). \tag{A.13}$$

Consider

$$\begin{aligned} A_n(\varepsilon) - A(\varepsilon) &= \big(\hat{F}_n(t_{(n)} - \varepsilon/2) - F(\tau_H - \varepsilon/2)\big)\big(\hat{F}_n(t_{(n)} - \varepsilon/2) + F(\tau_H - \varepsilon/2)\big) \\ &- \big(\hat{F}_n(t_{(n)}) - F(\tau_H)\big)\hat{F}_n(t_{(n)} - \varepsilon) - F(\tau_H)\big(\hat{F}_n(t_{(n)} - \varepsilon) - F(\tau_H - \varepsilon)\big) \\ &=: \mathbf{C}_n(\varepsilon)\mathbf{D}_n(\varepsilon), \end{aligned} \tag{A.14}$$

where[1]

$$\mathbf{C}_n(\varepsilon) = \big[(\hat{F}_n(t_{(n)} - \varepsilon/2) + F(\tau_H - \varepsilon/2), -\hat{F}_n(t_{(n)} - \varepsilon), -\hat{F}_n(t_{(n)}\big]$$

and

$$\mathbf{D}_n(\varepsilon) = \big[(\hat{F}_n(t_{(n)} - \varepsilon/2) - F(\tau_H - \varepsilon/2), \hat{F}_n(t_{(n)} - F(\tau_H), \hat{F}_n(t_{(n)} - \varepsilon) - F(\tau_H - \varepsilon)\big].$$

We relate the terms in $\mathbf{D}_n(\varepsilon)$ to $X_n(t)$. First,

$$\begin{aligned} &\sqrt{n}\big(\hat{F}_n(t_{(n)} - \varepsilon) - F(\tau_H - \varepsilon)\big) \\ &= \sqrt{n}\big(\hat{F}_n(t_{(n)} - \varepsilon) - F(t_{(n)} - \varepsilon)\big) + \sqrt{n}\big(F(t_{(n)} - \varepsilon) - F(\tau_H - \varepsilon)\big) \\ &= X_n(t_{(n)} - \varepsilon) + o_P(1). \end{aligned} \tag{A.15}$$

Here the $o_P(1)$ term comes about as follows. Since F is differentiable with density f, we have by Taylor expansion

$$\sqrt{n}\big(F(t_{(n)} - \varepsilon) - F(\tau_H - \varepsilon)\big) = \sqrt{n} f(\xi_n(\varepsilon))\big(t_{(n)} - \tau_H\big),$$

where $\xi_n(\varepsilon) \in [t_{(n)} - \varepsilon, \tau_H - \varepsilon] \xrightarrow{\text{P}} \tau_H - \varepsilon$ as $n \to \infty$. $f(\tau_h)$ is positive and for $\delta > 0$

$$\begin{aligned} P\big(\sqrt{n}(t_{(n)} - \tau_H) < \delta\big) &= P\big(\tau_H - \delta/\sqrt{n} < t_{(n)} < \tau_H\big) \\ &= 1 - H^n(\tau_H - \delta/\sqrt{n}) = 1 - e^{-n \log H(\tau_H - \delta/\sqrt{n})}. \end{aligned}$$

[1] We set vectors and matrices in boldface. Vectors are column vectors but to save space are written as rows with elements separated by commas.

Then, because of the assumption $n\overline{G}(\tau_H - \delta/\sqrt{n}) \to \infty$ and since $\overline{F}_0(\tau_H - \delta/\sqrt{n}) \to \overline{F}_0(\tau_H) > 0$, we have

$$\lim_{n\to\infty} n \log H(\tau_H - \delta/\sqrt{n}) = \lim_{n\to\infty} n\overline{H}(\tau_H - \delta/\sqrt{n}) = \lim_{n\to\infty} n\overline{F}_0(\tau_H - \delta/\sqrt{n})\overline{G}(\tau_H - \delta/\sqrt{n}) = \infty,$$

thus $\lim_{n\to\infty} P\big(\sqrt{n}(t_{(n)} - \tau_H) < \delta\big) = 1$ for all $\delta > 0$, and this shows that $\sqrt{n}(t_{(n)} - \tau_H) \xrightarrow{\text{P}} 0$, or, equivalently, $t_{(n)} = \tau_H + o_P(1/\sqrt{n})$. Thus (A.15) holds as claimed, and likewise we find

$$\sqrt{n}\big(\hat{F}_n(t_{(n)} - \varepsilon/2) - F(\tau_H - \varepsilon/2)\big) = X_n(t_{(n)} - \varepsilon/2) + o_P(1). \tag{A.16}$$

The remaining component of $\mathbf{D}_n(\varepsilon)$ is

$$\sqrt{n}\big(\hat{F}_n(t_{(n)}) - F(\tau_H)\big) = X_n(t_{(n)}) + o_P(1), \tag{A.17}$$

by (A.10). Thus we can write

$$\sqrt{n}\mathbf{D}_n(\varepsilon) = \big[X_n(t_{(n)} - \varepsilon/2), X_n(t_{(n)}), X_n(t_{(n)} - \varepsilon\big] + o_P(1),$$

and by the finite dimensional convergence of $(X_n(t))$ and the continuous mapping theorem this converges in distribution to the vector

$$\big[\overline{F}(\tau_H - \varepsilon/2)Z(\tau_H - \varepsilon/2), \overline{F}(\tau_H)Z(\tau_H), \overline{F}(\tau_H - \varepsilon)Z(\tau_H - \varepsilon)\big].$$

Letting $a(x) := \overline{F}(\tau_H - x)$, $x \in [0, \tau_H]$, we write the result as

$$\sqrt{n}\mathbf{D}_n(\varepsilon) \xrightarrow{\text{D}} \big[a(\varepsilon/2)Z(\tau_H - \varepsilon/2), a(0)Z(\tau_H), a(\varepsilon)Z(\tau_H - \varepsilon)\big]. \tag{A.18}$$

The vector on the RHS is trivariate normal $\mathbf{N}(\mathbf{0}, \mathbf{\Sigma}(\text{''})$, where

$$\mathbf{\Sigma}(\varepsilon) = \begin{bmatrix} a^2(\varepsilon/2)v(\tau_H - \varepsilon/2) & a(\varepsilon/2)a(0)v(\tau_H - \varepsilon/2) & a(\varepsilon/2)a(\varepsilon)v(\tau_H - \varepsilon) \\ * & a^2(0)v(\tau_H) & a(0)a(\varepsilon)v(\tau_H - \varepsilon) \\ * & * & a^2(\varepsilon)v(\tau_H - \varepsilon) \end{bmatrix} \tag{A.19}$$

(with the "*" elements filled in by symmetry). Having dealt with $A_n(\varepsilon)$ we look at

$$\begin{aligned} B_n(\varepsilon) - B(\varepsilon) &= 2\big(\hat{F}_n(t_{(n)} - \varepsilon/2) - F(\tau_H - \varepsilon/2)\big) \\ &\qquad - \big(\hat{F}_n(t_{(n)}) - F(\tau_H)\big) - \big(\hat{F}_n(t_{(n)} - \varepsilon/2) - F(\tau_H - \varepsilon/2)\big) \\ &=: [2,1,1]\mathbf{D}_n(\varepsilon). \end{aligned} \tag{A.20}$$

Putting together (A.14) and (A.20) gives

$$\begin{aligned} \sqrt{n}\big[A_n(\varepsilon) - A(\varepsilon), B_n(\varepsilon) - B(\varepsilon)\big] &= [\mathbf{C}_n(\varepsilon)\mathbf{D}_n(\varepsilon), [2,1,1]]\mathbf{D}_n(\varepsilon) + o_P(1) \\ &= \mathbf{J}(\varepsilon)\mathbf{D}_n(\varepsilon) + o_P(1), \end{aligned} \tag{A.21}$$

where $\mathbf{J}(\varepsilon)$ is the 2×3 matrix formed by stacking vector $\mathbf{C}(\varepsilon)$ on top of vector $[2,1,1]$.

The next step is to form

$$\begin{aligned} \sqrt{n}\big(\hat{p}_G(n,\varepsilon) - p_G(\varepsilon)\big) &= \sqrt{n}\Big(\frac{A_n(\varepsilon)}{B_n(\varepsilon)} - \frac{A(\varepsilon)}{B(\varepsilon)}\Big) \\ &= \sqrt{n}\Big(\frac{A_n(\varepsilon) - A(\varepsilon)}{B_n(\varepsilon)} - A(\varepsilon)\frac{B_n(\varepsilon) - B(\varepsilon)}{B_n(\varepsilon)B(\varepsilon)}\Big) \\ &= \frac{1}{B(\varepsilon)}\Big(\frac{B(\varepsilon)}{B_n(\varepsilon)}\Big)\Big[1, \frac{A(\varepsilon)}{B(\varepsilon)}\Big]^T \mathbf{J}(\varepsilon)\sqrt{n}\mathbf{D}_n(\varepsilon) \\ &:= \mathbf{K}(\varepsilon)\sqrt{n}\mathbf{D}_n(\varepsilon), \end{aligned} \tag{A.22}$$

from which we see, by (A.18), that $\sqrt{n}\big(\hat{p}_G(n,\varepsilon) - p_G(\varepsilon)\big)$ is asymptotically normal with mean zero and variance $\sigma^2(\varepsilon) := \mathbf{K}^T(\varepsilon)\mathbf{\Sigma}(\varepsilon)\mathbf{K}(\varepsilon)$ for the given $\varepsilon > 0$.

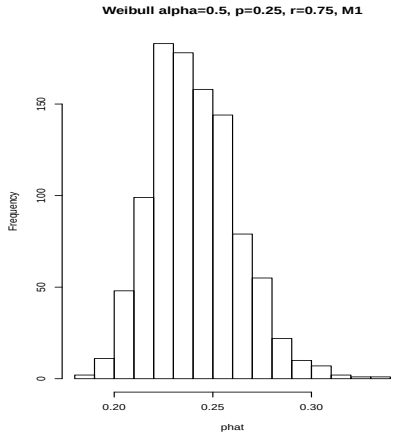
It follows from this analysis that

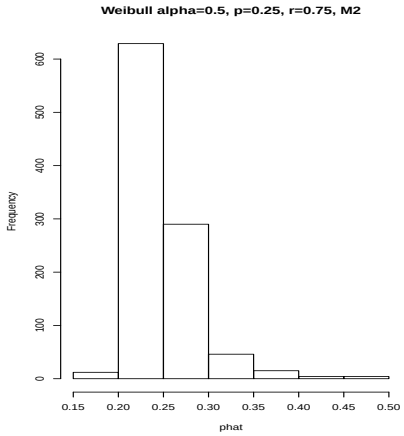
$$\frac{\sqrt{n}}{\sigma(\varepsilon)}\big(\hat{p}_G(n,\varepsilon) - p_G(\varepsilon)\big)$$

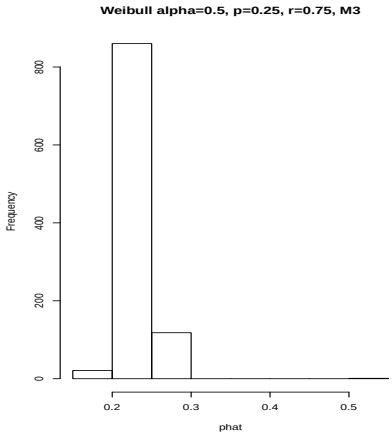
is asymptotically standard normally distributed when $n \to \infty$. For the second part of the proof, we want to let $\varepsilon \downarrow 0$. We can write down a reasonably explicit expression for the limit of $\sigma(\varepsilon)$ as $\varepsilon \downarrow 0$, implying that $\hat{p}_G(n,\varepsilon) - p_G(\varepsilon)$ is approximately normally distributed in large samples and for ε close to 0. This is the result we apply in practice: it allows us to claim that, for large sample size n and small ε, the estimator $\hat{p}_G(n,\varepsilon)$ is close to $p_G(\varepsilon)$, plus or minus 1.96 times a standard error. This then has the usual interpretation of a confidence interval. We do not attempt to calculate an explicit expression for the limit of $\sigma(\varepsilon)$ as it is really only of academic

interest. In a practical situation we obtain the standard error by bootstrapping or avoid it by studentising. Figures 6.1 and 6.2 illustrate the reasonableness of the normal approximation in a variety of situations.

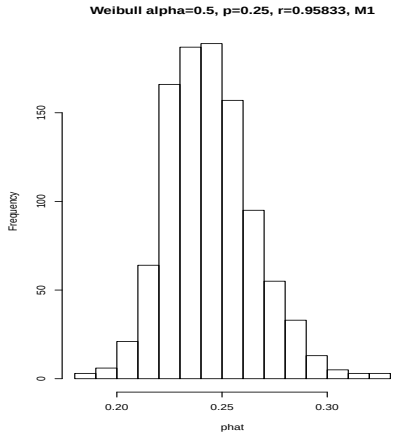
A.3 Plots and results for Section 6.4.3

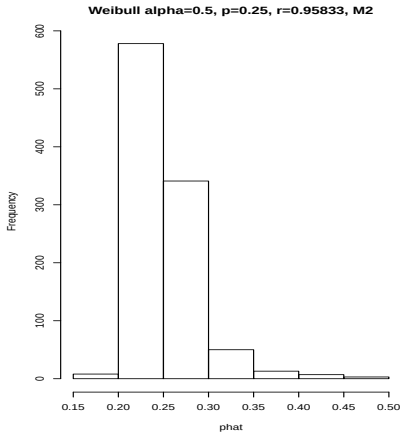


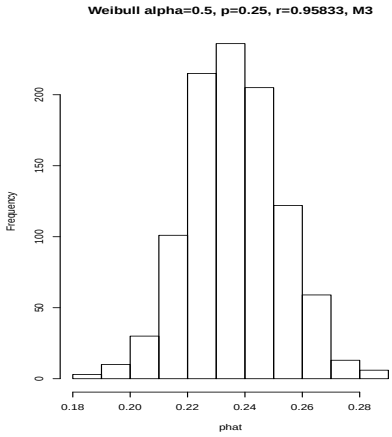




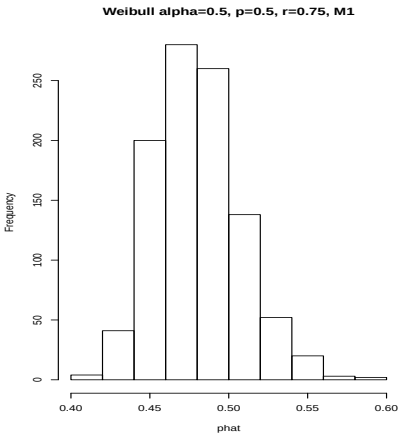
r=0.75	Method 1	Method 2	Method 3
MSE	0.0006	0.0013	0.0007
Mean	0.2410	0.2482	0.2316
Standard Err	0.0217	0.0365	0.0183

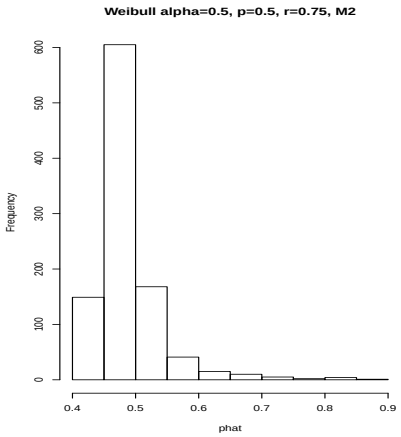


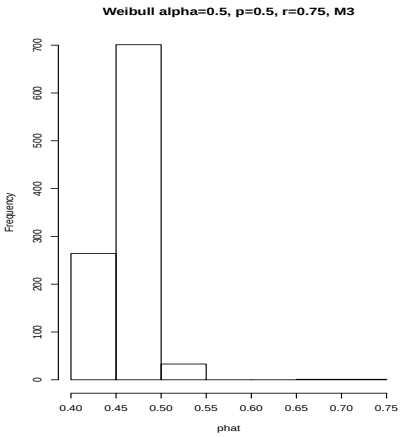




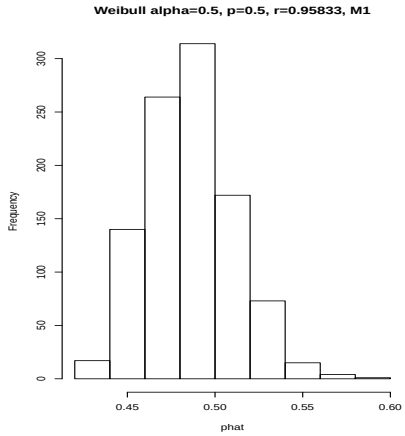
r=0.95	Method 1	Method 2	Method 3
MSE	0.0005	0.0012	0.0004
Mean	0.2444	0.2518	0.2366
Standard Err	0.0209	0.0346	0.0161

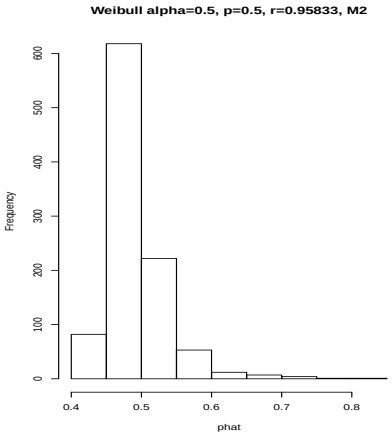


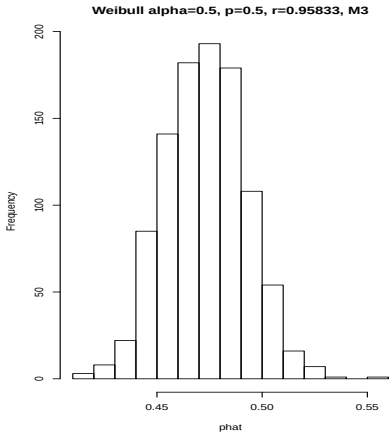




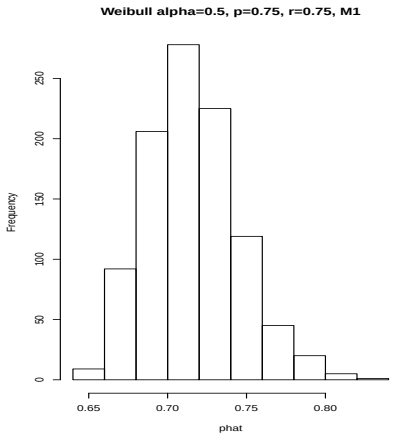
r=0.75	Method 1	Method 2	Method 3
MSE	0.0011	0.0029	0.0018
Mean	0.4802	0.4884	0.4633
Standard Err	0.0268	0.0523	0.0222

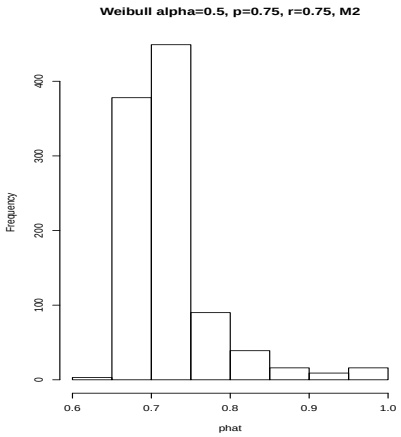


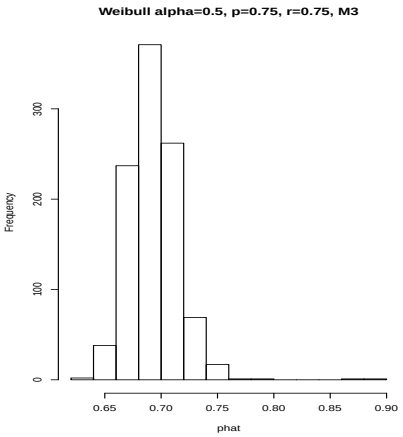




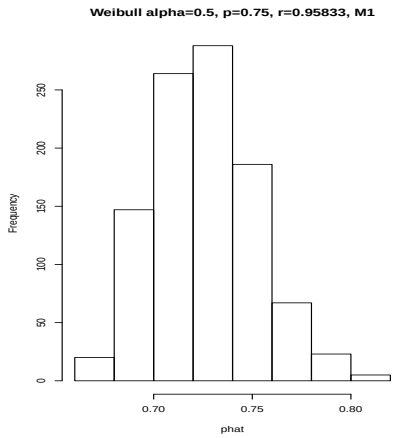
r=0.95	Method 1	Method 2	Method 3
MSE	0.0008	0.0019	0.0011
Mean	0.4858	0.4924	0.4730
Standard Err	0.0250	0.0427	0.0191

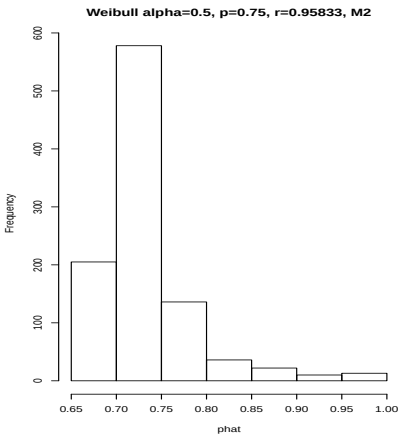


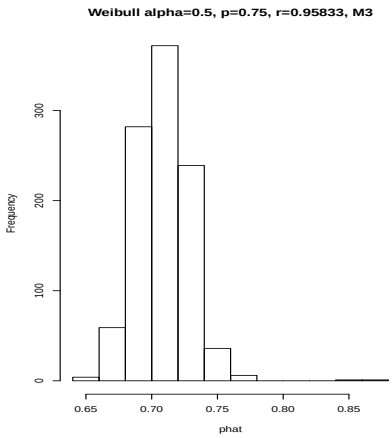




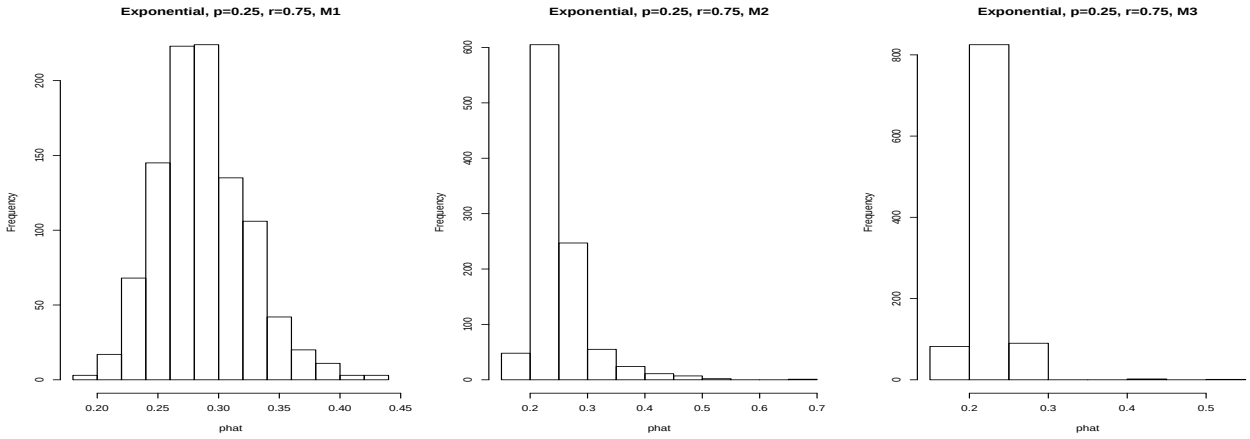
r=0.75	Method 1	Method 2	Method 3
MSE	0.0020	0.0039	0.0037
Mean	0.7158	0.7225	0.6932
Standard Err	0.0285	0.0559	0.0220



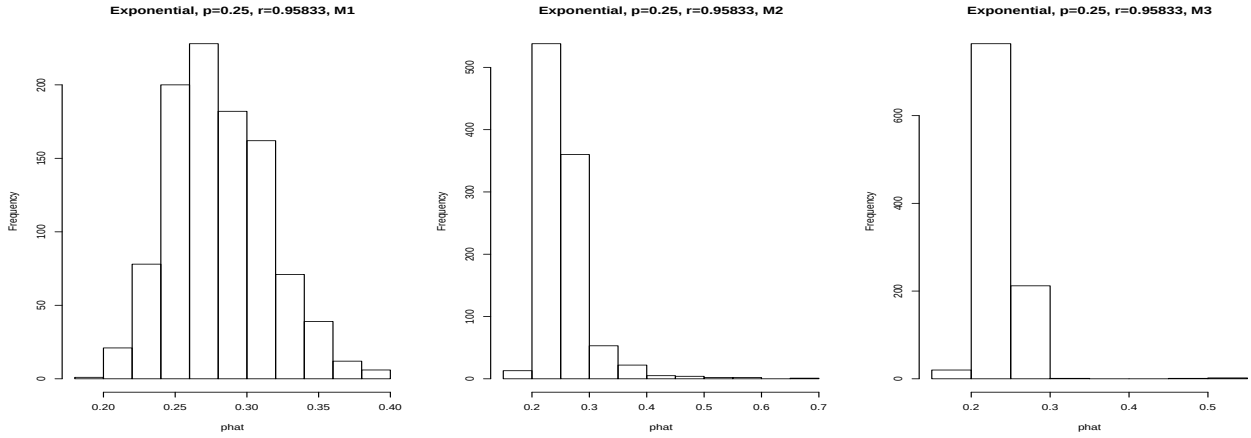




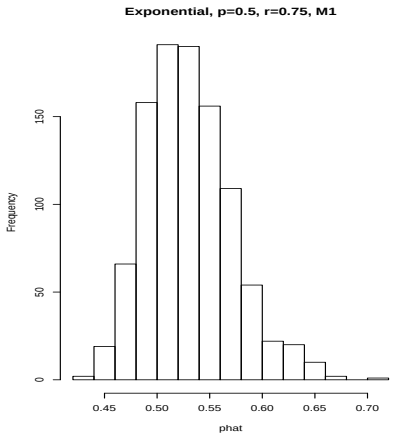
r=0.95	Method 1	Method 2	Method 3
MSE	0.0012	0.0029	0.0021
Mean	0.7256	0.7343	0.7088
Standard Err	0.0256	0.0516	0.0199

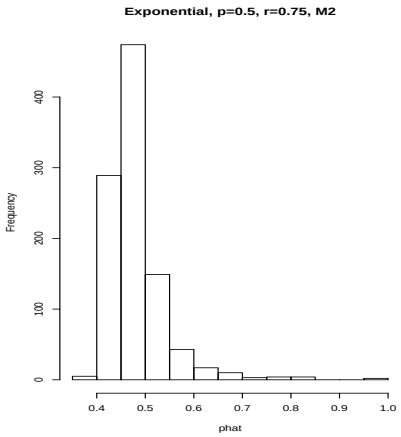


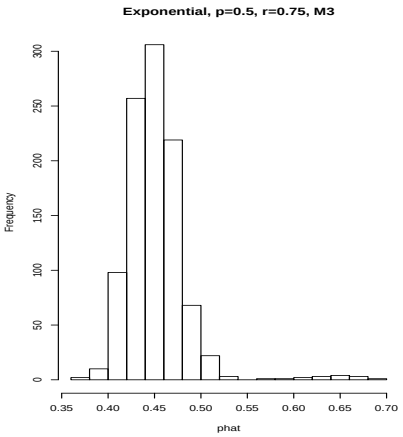
r=0.75	Method 1	Method 2	Method 3
MSE	0.0027	0.0024	0.0011
Mean	0.2867	0.2501	0.2259
Standard Err	0.0217	0.0365	0.0183



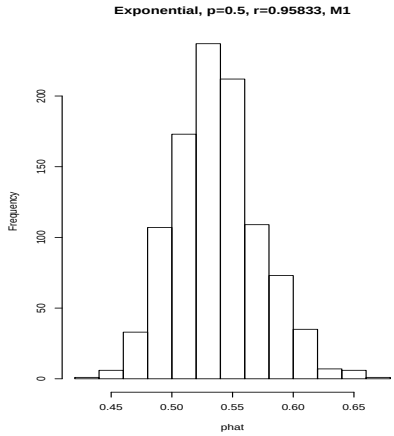
r=0.95	Method 1	Method 2	Method 3
MSE	0.0021	0.0020	0.0007
Mean	0.2807	0.2554	0.2359
Standard Err	0.0345	0.0444	0.0234

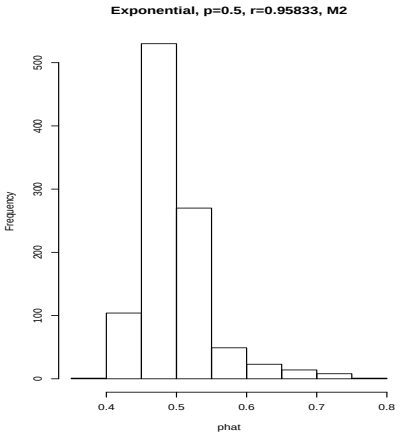


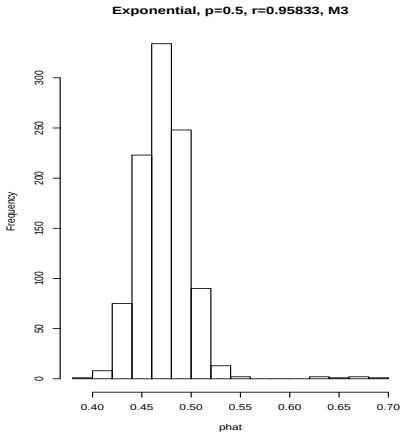




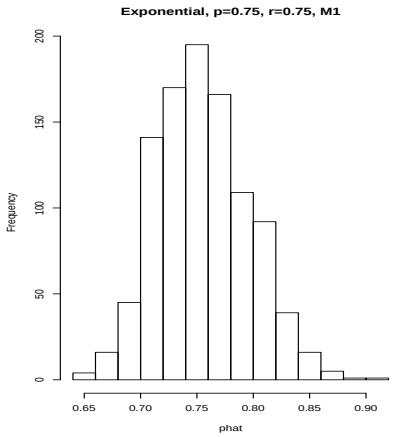
r=0.75	Method 1	Method 2	Method 3
MSE	0.0026	0.0039	0.0034
Mean	0.5306	0.4806	0.4519
Standard Err	0.0408	0.0595	0.0334

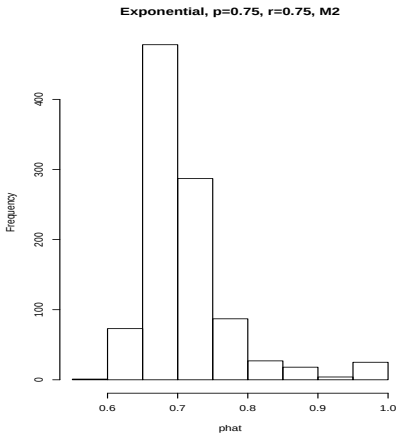


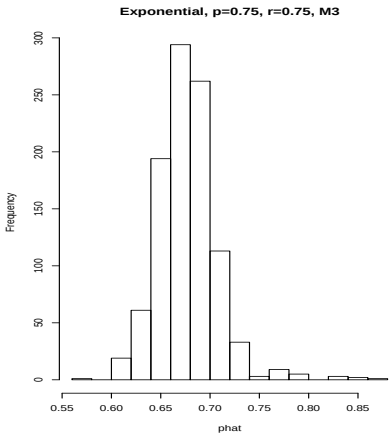




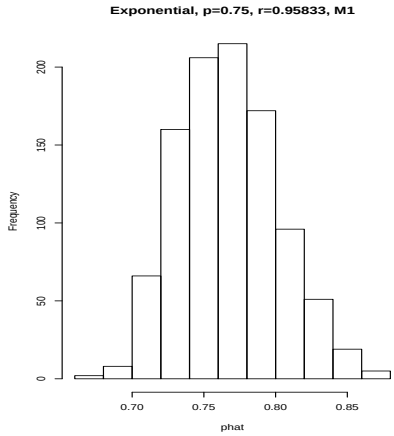
r=0.95	Method 1	Method 2	Method 3
MSE	0.0026	0.0024	0.0015
Mean	0.5369	0.4976	0.4726
Standard Err	0.0355	0.0489	0.0270

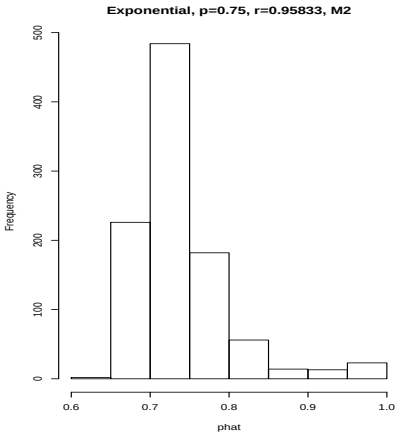


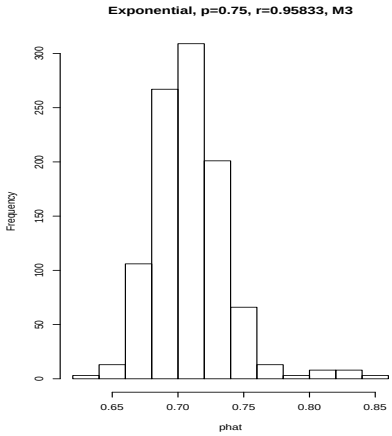




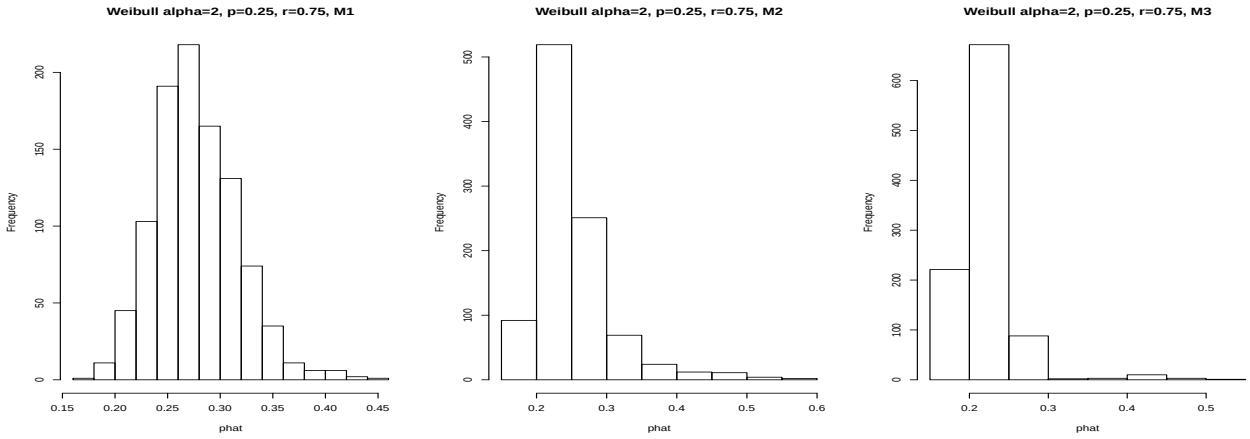
r=0.75	Method 1	Method 2	Method 3
MSE	0.0017	0.0058	0.0063
Mean	0.7556	0.7103	0.6768
Standard Err	0.0404	0.0652	0.0305



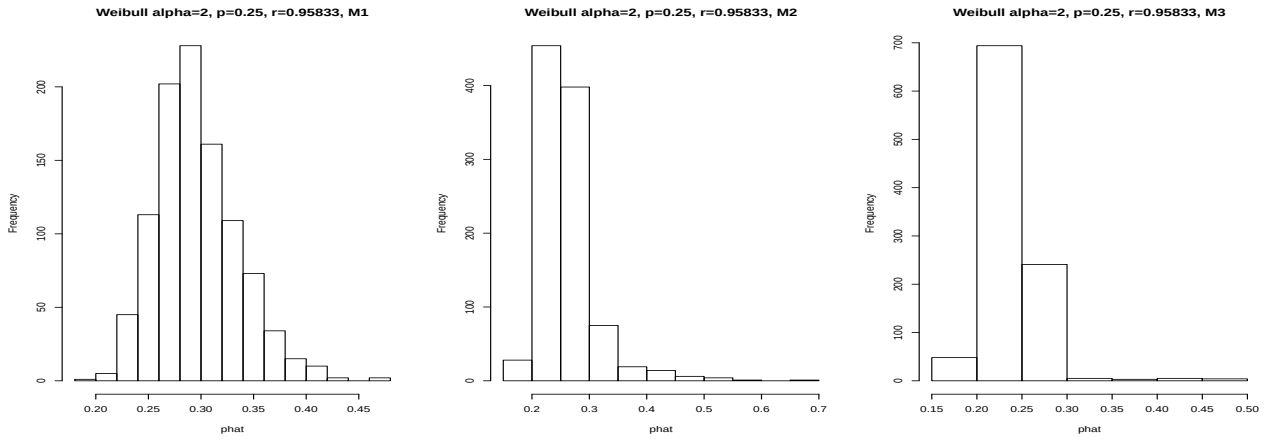




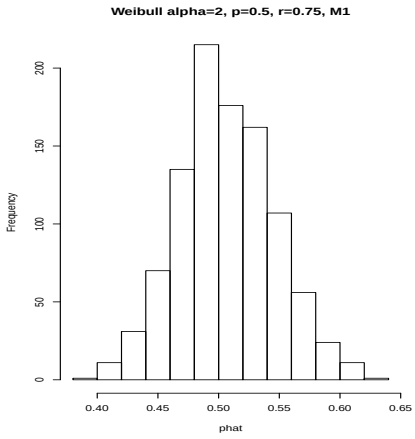
r=0.95	Method 1	Method 2	Method 3
MSE	0.0015	0.0037	0.0025
Mean	0.7670	0.7393	0.7093
Standard Err	0.0342	0.0598	0.0287

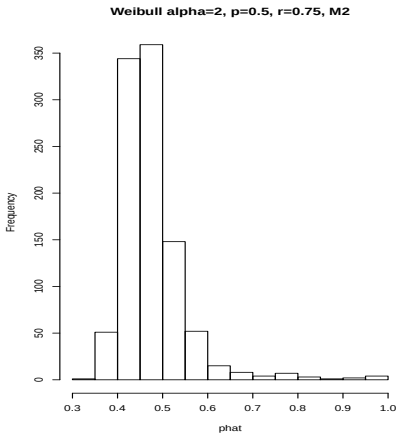


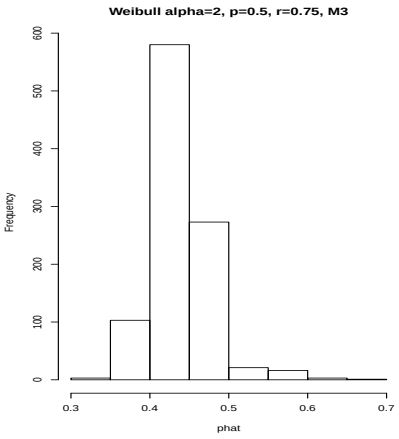
r=0.75	Method 1	Method 2	Method 3
MSE	0.0023	0.0031	0.0021
Mean	0.2772	0.2508	0.2217
Standard Err	0.0395	0.0559	0.0183



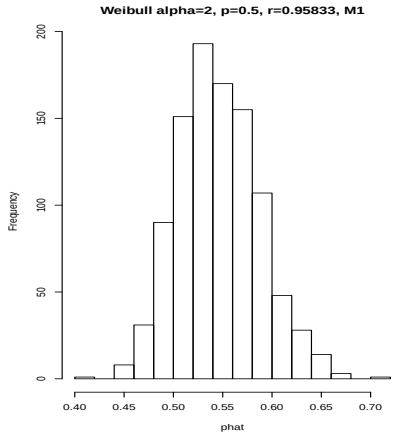
r=0.95	Method 1	Method 2	Method 3
MSE	0.0036	0.0025	0.0011
Mean	0.2962	0.2596	0.2377
Standard Err	0.0388	0.0488	0.0307

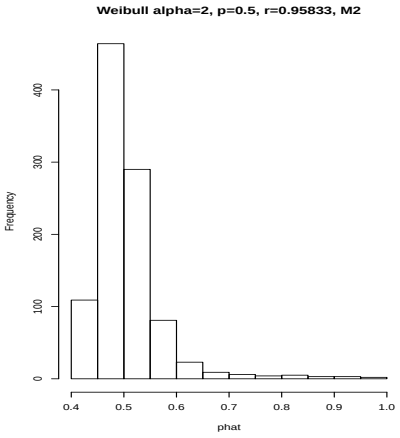


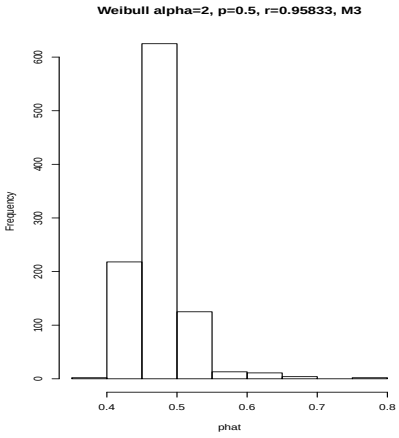




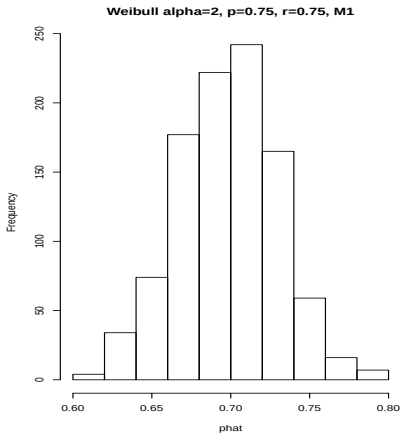
r=0.75	Method 1	Method 2	Method 3
MSE	0.0016	0.0062	0.0052
Mean	0.5063	0.4758	0.4383
Standard Err	0.0392	0.0747	0.0370

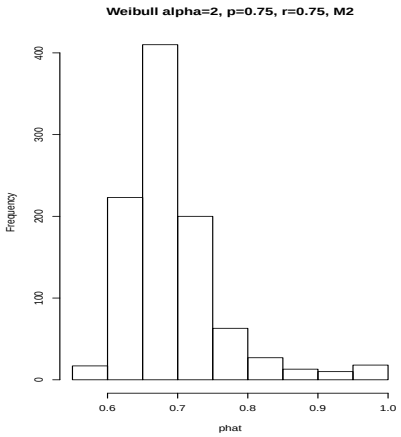


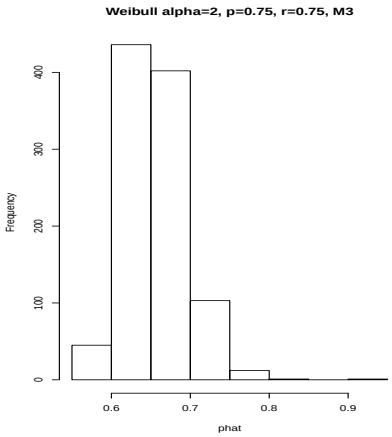




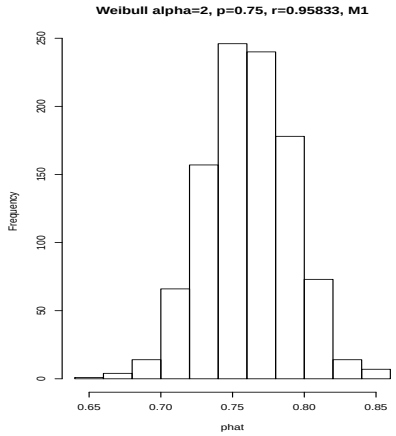
r=0.95	Method 1	Method 2	Method 3
MSE	0.0038	0.0046	0.0020
Mean	0.5456	0.5051	0.4747
Standard Err	0.0411	0.0673	0.0370

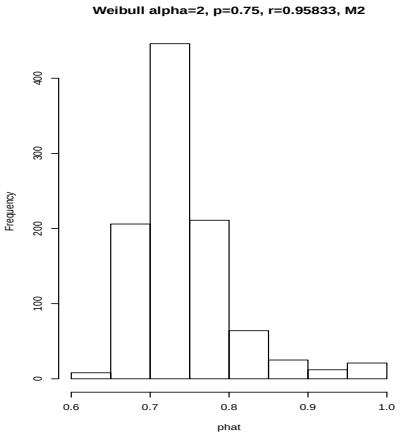


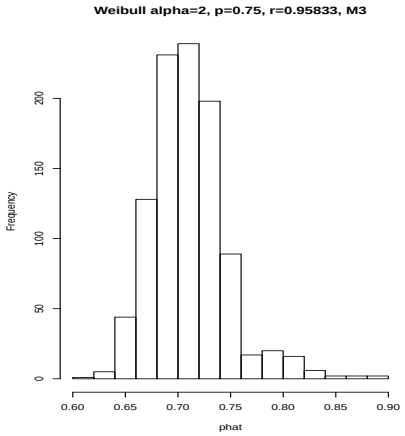




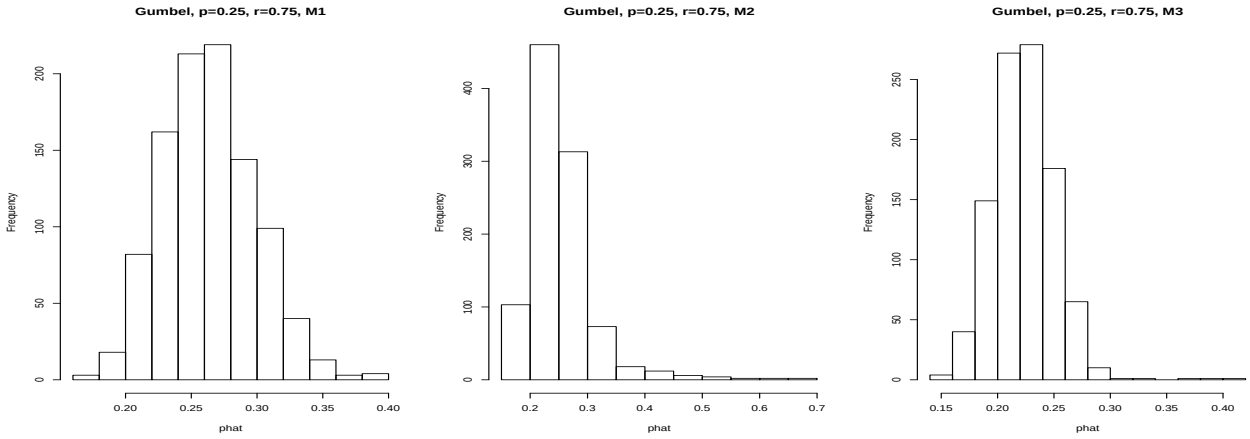
r=0.75	Method 1	Method 2	Method 3
MSE	0.0037	0.0081	0.0103
Mean	0.6979	0.6930	0.6557
Standard Err	0.0315	0.0700	0.0374



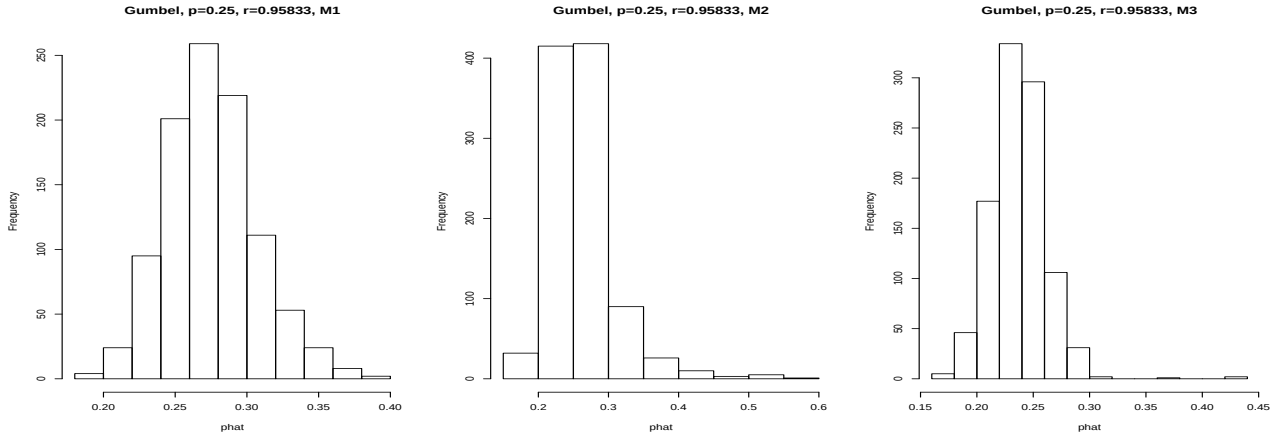




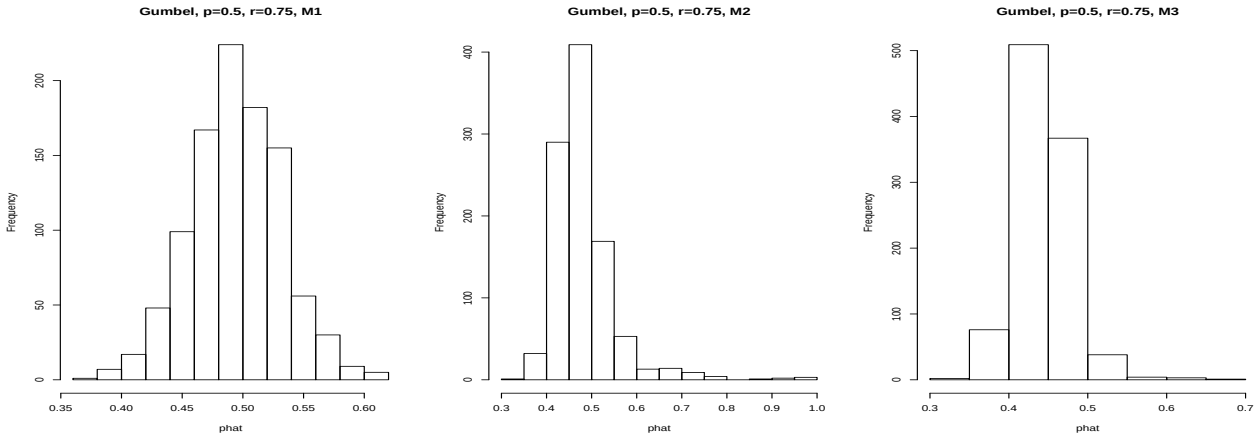
r=0.95	Method 1	Method 2	Method 3
MSE	0.0010	0.0038	0.0028
Mean	0.7612	0.7427	0.7104
Standard Err	0.0302	0.0613	0.0358



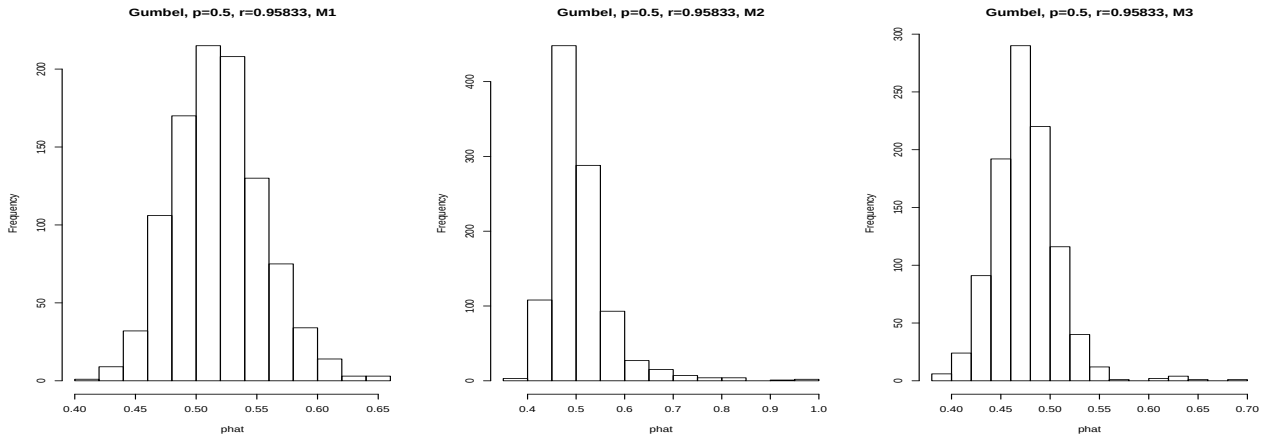
r=0.75	Method 1	Method 2	Method 3
MSE	0.0015	0.0035	0.0014
Mean	0.2638	0.2530	0.2232
Standard Err	0.0355	0.0590	0.0271



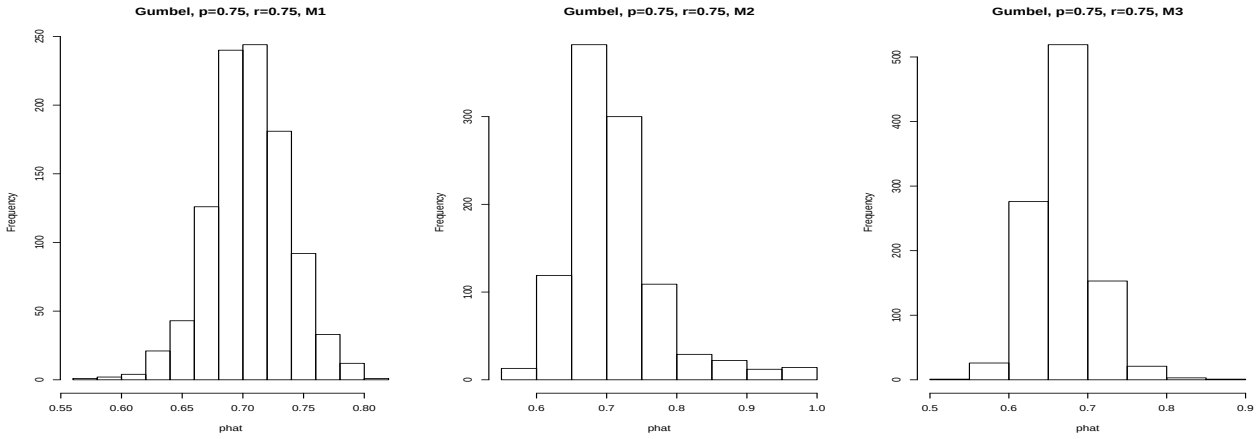
r=0.95	Method 1	Method 2	Method 3
MSE	0.0016	0.0024	0.0008
Mean	0.2753	0.2622	0.2371
Standard Err	0.0314	0.0474	0.0243



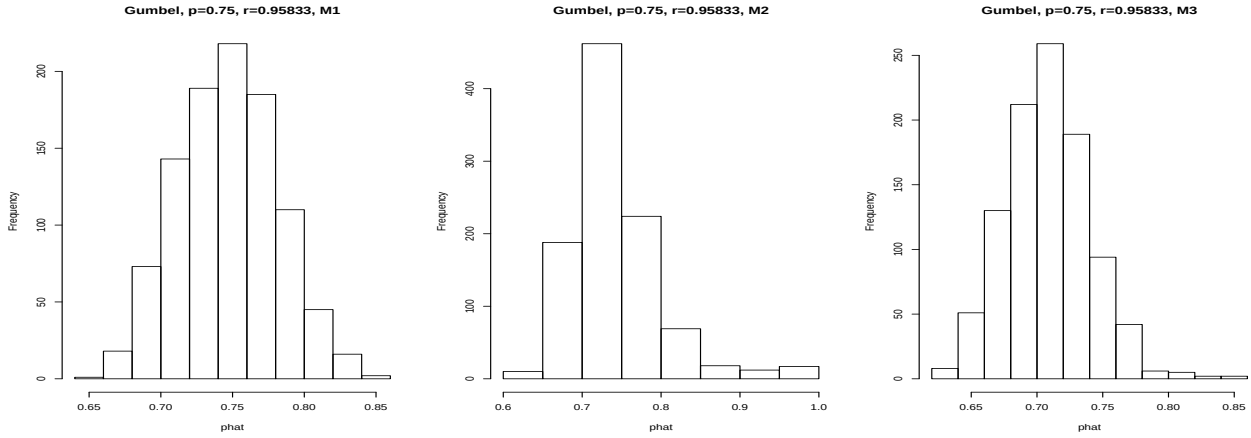
r=0.75	Method 1	Method 2	Method 3
MSE	0.0014	0.0052	0.0041
Mean	0.4948	0.4818	0.4454
Standard Err	0.0373	0.0697	0.0341



r=0.95	Method 1	Method 2	Method 3
MSE	0.0017	0.0041	0.0017
Mean	0.5185	0.5057	0.4741
Standard Err	0.0367	0.0636	0.0322



r=0.75	Method 1	Method 2	Method 3
MSE	0.0030	0.0061	0.0079
Mean	0.7056	0.7088	0.6694
Standard Err	0.0325	0.0667	0.0371



r=0.95	Method 1	Method 2	Method 3
MSE	0.0012	0.0034	0.0028
Mean	0.7468	0.7421	0.7085
Standard Err	0.0343	0.0581	0.0323

A.4 Gumbel and Weibull fit to the breast cancer data

The Gumbel cumulative distribution function we use is

$$F_g(t) = \exp\left(-\exp\left(-\frac{t-\mu_g}{\sigma_g}\right)\right),$$

having location and scale parameters μ_g and σ_g, and density function

$$f_g(t;\mu_g,\sigma_g) = \frac{1}{\sigma_g}\exp\left(-\frac{t-\mu_g}{\sigma_g}\right)\exp\left(-\exp\left(-\frac{t-\mu_g}{\sigma_g}\right)\right).$$

The support for the distribution is $\mathbb{R}$, so we need to truncate it before applying it in a survival context. The truncated Gumbel distribution has density function

$$f_{Tg}(t;\mu_g,\sigma_g) = f_g(t|T \geq 0) = \frac{f_g(t)}{P(T \geq 0)} = \frac{f_g(t;\mu_g,\sigma_g)}{\overline{F}_g(0;\mu_g,\sigma_g)},\ t \geq 0,$$

with corresponding survival function

$$S_{Tg}(t;\mu_g,\sigma_g) = 1 - P(T < t|T \geq 0) = 1 - \frac{F_g(t;\mu_g,\sigma_g) - F_g(0;\mu_g,\sigma_g)}{\overline{F}_g(0;\mu_g,\sigma_g)}.$$

Using a mixture cure model structure as defined in Definition 2.1, the likelihood function for the Gumbel mixture model is

$$L(\mu_g,\sigma_g,p) = \prod_{i=1}^{n}\left[pf_{Tg}(t_i)\right]^{c_i}\left[1-p+pS_{Tg}(t_i)\right]^{1-c_i}. \tag{A.23}$$

The Gumbel model fits to the breast cancer data are shown in Figure A.6. Like the Weibull, the Gumbel is a very good fit for the major parts of the distributions, for each cancer stage, though a close inspection shows a slightly worse fit for the Gumbel than for the Weibull. Again there are some noticeable departures at the extreme right-hand ends of the distributions.

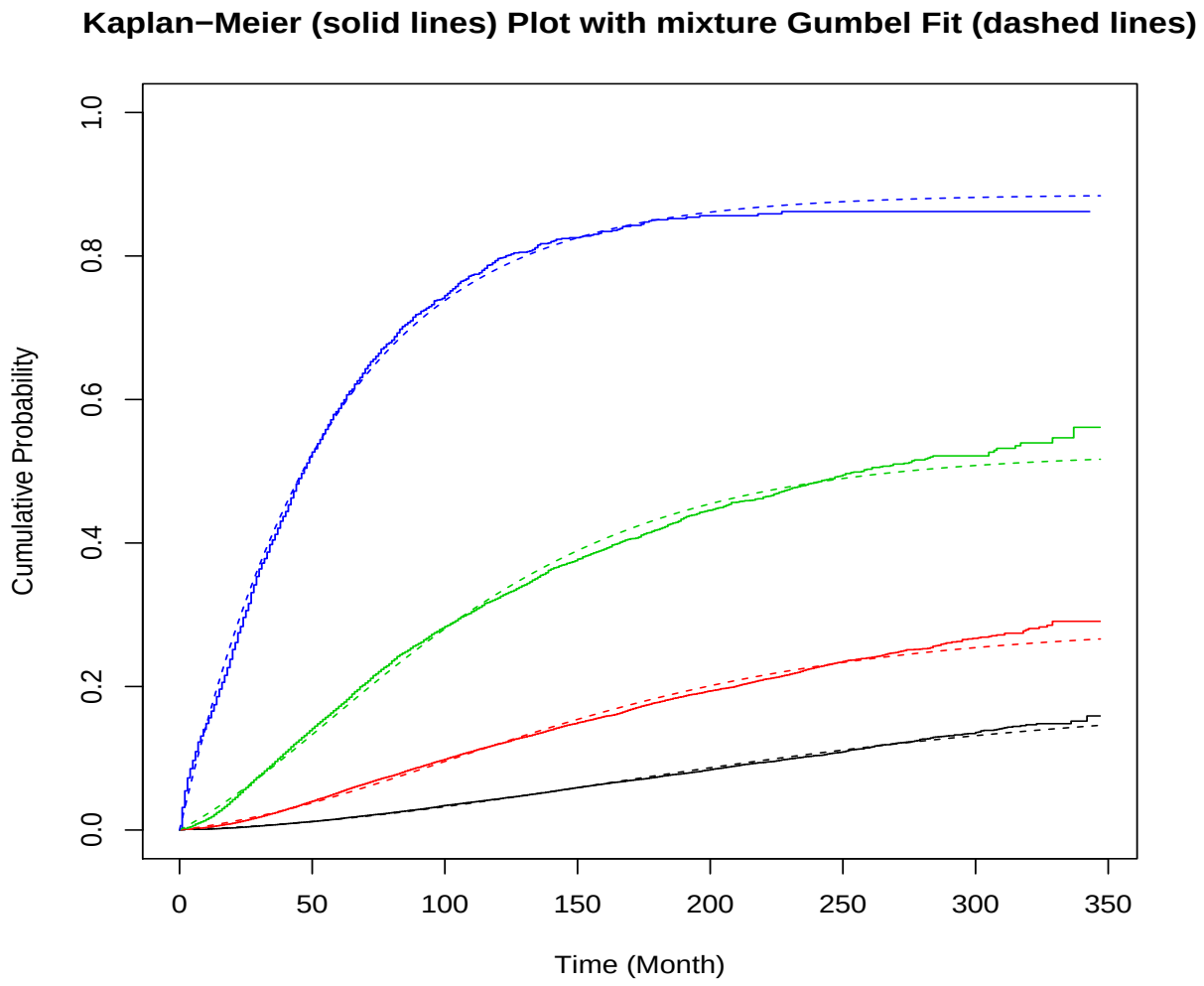


FIGURE A.6: Breast cancer data, Kaplan-Meier plots with staging. Black, Red, Green, Blue = Stage I, II, III, IV. Full lines = fitted Gumbel mixture models.

The detailed parameterisation is given as in Table A.2.

	$1-p$	μ	σ
Stage I	0.8178 (0.0109)	158.3656 (7.0957)	122.9394 (9.2741)
Stage II	0.7152 (0.0087)	100.2017 (2.5766)	89.7441 (3.6690)
Stage III	0.4740 (0.0102)	57.1136 (1.6641)	69.9008 (2.8800)
Stage IV	0.1143 (0.0178)	-906.4837 (531.5127)	55.8485 (5.4636)

TABLE A.2: Gumbel Parameterization. Bootstrap standard errors are in brackets, N_B=200.

A mixture Weibull cure model, whose details can be found in Section 2.2.2, was also fitted

to the data, the resulting plot was presented in Section 6.6 (Figure 6.3). The detailed parameterisation is given as in Table A.3.

	$1-p$	α	λ
Stage I	0.7372 (0.0351)	1.5255 (0.0287)	370.1728 (40.9734)
Stage II	0.6903 (0.0103)	1.4674 (0.0196)	196.3889 (7.0215)
Stage III	0.4631 (0.0100)	1.3447 (0.0178)	125.6206 (2.9894)
Stage IV	0.1175 (0.0104)	1.0133 (0.0168)	55.3175 (1.5428)

TABLE A.3: Weibull Parameterization. Bootstrap standard errors are in brackets, N_B=200.

A.5 Simulation results with different r

$r = 1.5$

		Cens. prop. (%)	$\hat{\tau}_G$	$\hat{\tau}_{F_0}$
$p = 0.25$	Gumbel	86.1918	7.6138	10.4320
	Weibull ($\alpha = 0.5$)	78.1607	13.3217	56.8599
	Exponential	80.4158	4.3341	7.5300
	Weibull ($\alpha = 2$)	84.1252	2.3209	2.7333
$p = 0.5$	Gumbel	72.3366	7.6164	10.5275
	Weibull ($\alpha = 0.5$)	56.2536	13.3606	57.4578
	Exponential	60.8454	4.3285	7.4786
	Weibull ($\alpha = 2$)	68.1254	2.3223	2.7283
$p = 0.75$	Gumbel	58.4688	7.613	10.5108
	Weibull ($\alpha = 0.5$)	34.5278	13.4190	57.8423
	Exponential	41.2609	4.3332	7.5128
	Weibull ($\alpha = 2$)	52.1773	2.3241	2.7317

TABLE A.4: Simulated Data Summary when $r = 1.5$.

		Gumbel	Weibull ($\alpha = 0.5$)	Exponential	Weibull ($\alpha = 2$)
$\hat{p}_G(n, \varepsilon)$	MSE	0.00132	0.00034	0.00058	0.00184
	Mean	0.27787	0.24810	0.25996	0.28588
	Standard Err	0.02330	0.01847	0.02189	0.02352
$\hat{p}_n$	MSE	0.00040	0.00028	0.00028	0.00036
	Mean	0.24702	0.24418	0.24703	0.24768
	Standard Err	0.01976	0.01574	0.01643	0.01884
$\hat{p}_w$	MSE	0.00037	0.00029	0.00031	0.00041
	Mean	0.24630	0.25070	0.25084	0.25168
	Standard Err	0.01900	0.01692	0.01773	0.02022

TABLE A.5: Simulation Summary Statistics with $r = 1.5$ and $p = 0.25$.

		Gumbel	Weibull ($\alpha = 0.5$)	Exponential	Weibull ($\alpha = 2$)
$\hat{p}_G(n, \varepsilon)$	MSE	0.00147	0.00048	0.00099	0.00201
	Mean	0.52699	0.49447	0.51654	0.53746
	Standard Err	0.02716	0.02124	0.02674	0.02474
$\hat{p}_n$	MSE	0.00060	0.00051	0.00043	0.00051
	Mean	0.49406	0.48820	0.49413	0.49856
	Standard Err	0.02380	0.01926	0.01992	0.02252
$\hat{p}_w$	MSE	0.00066	0.00050	0.00045	0.00055
	Mean	0.48981	0.50149	0.50035	0.50025
	Standard Err	0.02361	0.02231	0.02116	0.02339

TABLE A.6: Simulation Summary Statistics with $r = 1.5$ and $p = 0.5$.

		Gumbel	Weibull ($\alpha = 0.5$)	Exponential	Weibull ($\alpha = 2$)
$\hat{p}_G(n, \varepsilon)$	MSE	0.00088	0.00059	0.00101	0.00120
	Mean	0.76632	0.74098	0.77022	0.77618
	Standard Err	0.02471	0.02256	0.02450	0.02277
$\hat{p}_n$	MSE	0.00059	0.00071	0.00050	0.00052
	Mean	0.74318	0.73068	0.74004	0.74720
	Standard Err	0.02332	0.01827	0.02003	0.02272
$\hat{p}_w$	MSE	0.00080	0.00042	0.00040	0.00053
	Mean	0.73220	0.75233	0.75155	0.75021
	Standard Err	0.02198	0.02029	0.02007	0.02295

TABLE A.7: Simulation Summary Statistics with $r = 1.5$ and $p = 0.75$.

$r = 2$

		Cens. prop. (%)	$\hat{\tau}_G$	$\hat{\tau}_{F_0}$
$p = 0.25$	Gumbel	84.1341	9.2612	10.4936
	Weibull ($\alpha = 0.5$)	77.4845	17.8733	56.7441
	Exponential	79.1769	5.6859	7.4700
	Weibull ($\alpha = 2$)	82.2756	2.9180	2.7260
$p = 0.5$	Gumbel	68.5198	9.2398	10.6078
	Weibull ($\alpha = 0.5$)	54.9487	17.7848	57.8635
	Exponential	58.3059	5.6803	7.4890
	Weibull ($\alpha = 2$)	64.3941	2.9187	2.7352
$p = 0.75$	Gumbel	52.5431	9.2648	10.5098
	Weibull ($\alpha = 0.5$)	32.4254	17.8169	58.2600
	Exponential	37.5037	5.6863	7.5291
	Weibull ($\alpha = 2$)	46.6445	2.9158	2.7253

TABLE A.8: Simulated Data Summary when $r = 2$.

		Gumbel	Weibull ($\alpha = 0.5$)	Exponential	Weibull ($\alpha = 2$)
$\hat{p}_G(n, \varepsilon)$	MSE	0.00091	0.00028	0.00031	0.00132
	Mean	0.27278	0.24881	0.25511	0.28136
	Standard Err	0.01985	0.01664	0.01678	0.01835
$\hat{p}_n$	MSE	0.00030	0.00024	0.00024	0.00031
	Mean	0.25008	0.24639	0.24919	0.24925
	Standard Err	0.01745	0.01504	0.01547	0.01761
$\hat{p}_w$	MSE	0.00035	0.00025	0.00026	0.00030
	Mean	0.24785	0.25059	0.25048	0.24940
	Standard Err	0.01861	0.01576	0.01605	0.01745

TABLE A.9: Simulation Summary Statistics with $r = 2$ and $p = 0.25$.

		Gumbel	Weibull ($\alpha = 0.5$)	Exponential	Weibull ($\alpha = 2$)
$\hat{p}_G(n, \varepsilon)$	MSE	0.00109	0.00040	0.00045	0.00141
	Mean	0.52492	0.49715	0.50602	0.53156
	Standard Err	0.02172	0.01988	0.02035	0.02040
$\hat{p}_n$	MSE	0.00049	0.00038	0.00035	0.00042
	Mean	0.49774	0.49341	0.49907	0.50029
	Standard Err	0.02194	0.01825	0.01867	0.02048
$\hat{p}_w$	MSE	0.00048	0.00037	0.00032	0.00043
	Mean	0.49634	0.50130	0.50038	0.50119
	Standard Err	0.02156	0.01931	0.01801	0.02076

TABLE A.10: Simulation Summary Statistics with $r = 2$ and $p = 0.5$.

		Gumbel	Weibull ($\alpha = 0.5$)	Exponential	Weibull ($\alpha = 2$)
$\hat{p}_G(n, \varepsilon)$	MSE	0.00060	0.00038	0.00044	0.00075
	Mean	0.76514	0.74492	0.75791	0.77107
	Standard Err	0.01924	0.01892	0.01956	0.01747
$\hat{p}_n$	MSE	0.00039	0.00040	0.00035	0.00036
	Mean	0.74811	0.73961	0.74723	0.75018
	Standard Err	0.01966	0.01709	0.01844	0.01902
$\hat{p}_w$	MSE	0.00042	0.00033	0.00031	0.00037
	Mean	0.74339	0.75205	0.75025	0.74887
	Standard Err	0.01952	0.01811	0.01748	0.01913

TABLE A.11: Simulation Summary Statistics with $r = 2$ and $p = 0.75$.

Bibliography

Ahuja, J., & Nash, S. W. (1967). The generalized gompertz-verhulst family of distributions. *Sankhyā: The Indian Journal of Statistics, Series A*, 141–156.

Akl, E. A., Briel, M., You, J. J., Sun, X., Johnston, B. C., Busse, J. W., Mulla, S., Lamontagne, F., Bassler, D., Vera, C., Alshurafa, M., Katsios, C. M., Zhou, Q., Cukierman-Yaffe, T., Gangji, A., Mills, E. J., Walter, S. D., Cook, D. J., Schünemann, H. J., … Guyatt, G. H. (2012). Potential impact on estimated treatment effects of information lost to follow-up in randomised controlled trials (lost-it): Systematic review. *Bmj*, *344*.

Altman, D., De Stavola, B., Love, S., & Stepniewska, K. (1995). Review of survival analyses published in cancer journals. *British journal of cancer*, *72*(2), 511–518.

Amdahl, J. (n.d.). Flexsurvcure: Flexible parametric cure models, 2019. *URL https://CRAN. R-project. org/package= flexsurvcure. R package version*, *1*(0), 117.

Ang, K. J. M. (2019). Conditional hypothesis testing. *Available at SSRN 3720414*.

Bartlett, M. S., & Kendall, D. (1946). The statistical analysis of variance-heterogeneity and the logarithmic transformation. *Supplement to the Journal of the Royal Statistical Society*, *8*(1), 128–138.

Berkson, J., & Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, *47*(259), 501–515.

Berliner, L. M., & Hill, B. M. (1988). Bayesian nonparametric survival analysis. *Journal of the American Statistical Association*, *83*(403), 772–779.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, *11*(1), 15–53.

Breslow, N. E. (1972). Contribution to discussion of paper by dr cox. *J. Roy. Statist. Soc., Ser. B*, *34*, 216–217.

Broadhurst, R. G., & Maller, R. A. (1992). The recidivism of sex offenders in the western australian prison population. *The British Journal of Criminology*, *32*(1), 54–80.

Buckley, J., & James, I. (1979). Linear regression with censored data. *Biometrika*, *66*(3), 429–436.

Burr, I. W. (1942). Cumulative frequency functions. *The Annals of mathematical statistics*, *13*(2), 215–232.

Cai, C., Zou, Y., Peng, Y., & Zhang, J. (2012). Smcure: An r-package for estimating semiparametric mixture cure models. *Computer methods and programs in biomedicine*, *108*(3), 1255–1260.

Chandrasekaran, N., Scharifker, D., Varsegi, G., & Almeida, Z. (2016). Breast metastasis from vaginal cancer. *Case Reports*, *2016*, bcr2016215895.

Chen, M.-H., Ibrahim, J. G., & Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, *94*(447), 909–919.

Chen, Y., Moustaki, I., & Zhang, H. (2020). A note on likelihood ratio tests for models with latent variables. *psychometrika*, *85*(4), 996–1012.

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003). Survival analysis part i: Basic concepts and first analyses. *British journal of cancer*, *89*(2), 232–238.

Collett, D. (2003). *Modelling survival data in medical research (2nd edition)*. Chapman; Hall/CRC.

Contant Jr, C. F. (1988). *Effects of informative censoring on the hazard function: Application to the proportional hazards regression model*. The University of Texas School of Public Health.

Cox, C. (2008). The generalized f distribution: An umbrella for parametric survival analysis. *Statistics in Medicine*, *27*(21), 4301–4312.

Cox, D. R. (1959). The analysis of exponentially distributed life-times with two types of failure. *Journal of the Royal Statistical Society: Series B (Methodological)*, *21*(2), 411–421.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, *62*(2), 269–276.

Cox, D. R., & Hinkley, D. V. (1979). *Theoretical statistics*. CRC Press.

de Haan, L. (1970). *On regular variation and its application to the weak convergence of sample extremes, by l. de haan*.

de Haan, L., Ferreira, A., & Ferreira, A. (2006). *Extreme value theory: An introduction* (Vol. 21). Springer.

do Nascimento, F. F., Gamerman, D., & Lopes, H. F. (2012). A semiparametric bayesian approach to extreme value estimation. *Statistics and Computing*, *22*(2), 661–675.

Donegan, W. L., & Redlich, P. N. (1996). Breast cancer in men. *Surgical Clinics of North America*, *76*(2), 343–363.

Douglas, S., & Hariharan, G. (1994). The hazard of starting smoking: Estimates from a split population duration model. *Journal of Health Economics*, *13*(2), 213–230.

Dupuy, J.-F. (2014). Accelerated failure time models: A review. *International Journal of Performability Engineering*, *10*(1).

Dyer, D. (1982). The convolution of generalized f distributions. *Journal of the American Statistical Association*, *77*(377), 184–189.

Elandt-Johnson, R. C., & Johnson, N. L. (1980). *Survival models and data analysis* (Vol. 110). John Wiley & Sons.

Escobar-Bach, M., & Keilegom, I. V. (2019). Non-parametric cure rate estimation under insufficient follow-up by using extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *81*(5), 861–880.

Escobar-Bach, M., Maller, R., Van Keilegom, I., & Zhao, M. (2020). Estimation of the cure rate for distributions in the gumbel maximum domain of attraction under insufficient follow-up. *Biometrika*.

Escobar-Bach, M., & Van Keilegom, I. (2019a). Nonparametric estimation of conditional cure models for heavy-tailed distributions and under insufficient follow-up. *arXiv preprint arXiv:1909.08436*.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 1041–1046.

Gamel, J. W., McLean, I. W., & Rosenberg, S. H. (1990). Proportion cured and mean log survival time as functions of tumour size. *Statistics in Medicine*, *9*(8), 999–1006.

Garthwaite, P. H., Jolliffe, I. T., Jolliffe, I., & Jones, B. (2002). *Statistical inference*. Oxford University Press on Demand.

Geerdens, C., Janssen, P., & Van Keilegom, I. (2020). Goodness-of-fit test for a parametric survival function with cure fraction. *Test*, *29*(3), 768–792.

Ghitany, M., Maller, R. A., & Zhou, S. (1994). Exponential mixture models with long-term survivors and covariates. *Journal of Multivariate Analysis*, *49*(2), 218–241.

Gumbel, E. J. (1944). Ranges and midranges. *The Annals of Mathematical Statistics*, *15*(4), 414–422.

Gupta, R. D., & Kundu, D. (2010). Generalized logistic distributions. *Journal of Applied Statistical Science*, *18*(1), 51.

Hall, P., & Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, 757–762.

Huang, G.-H. (2005). Model identifiability. *Encyclopedia of Statistics in Behavioral Science*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, *81*(348), 158–171.

Johnson, N. L. (1970). *Continuous univariate distributions* (tech. rep.).

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions, volume 2* (Vol. 289). John wiley & sons.

Kalbfleisch, J. D., & Prentice, R. L. (1980). *The statistical analysis of failure time data (1st edition)* (Vol. 360). John Wiley & Sons.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457–481.

Klebanov, L. B., & Yakovlev, A. Y. (2007). A new approach to testing for sufficient follow-up in cure-rate analysis. *Journal of Statistical Planning and Inference*, *137*(11), 3557–3569.

Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G., & Scheike, T. H. (2014). *Handbook of survival analysis*. CRC Press Boca Raton, FL:

Korn, E. L. (1986). Censoring distributions as a measure of follow-up in survival analysis. *Statistics in medicine*, *5*(3), 255–260.

Kuk, A. Y., & Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, *79*(3), 531–541.

Lagakos, S. W. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics*, 139–156.

Lawless, J. F. (1982). *Statistical models and methods for lifetime data*. John Wiley & Sons.

Li, C.-S., & Taylor, J. M. (2002). A semi-parametric accelerated failure time cure model. *Statistics in Medicine*, *21*(21), 3235–3247.

Li, C.-S., Taylor, J. M., & Sy, J. P. (2001). Identifiability of cure models. *Statistics & Probability Letters*, *54*(4), 389–395.

Lin, D. (2007). On the breslow estimator. *Lifetime Data Analysis*, *13*(4), 471–480.

López Segovia, L. (2014). Survival data analysis with heavy-censoring and long-term survivors. *Universitat Politècnica de Catalunya*.

Lu, W. (2010). Efficient estimation for an accelerated failure time model with a cure fraction. *Statistica Sinica*, *20*, 661–674.

Maller, R., & Resnick, S. (2021). Extremes of censored and uncensored lifetimes in survival data. *Extremes*, 1–31.

Maller, R., Resnick, S., & Shemehsavar, S. (2021a). Exact and asymptotic tests for sufficient followup in censored survival data. *arXiv preprint arXiv:2109.02817*.

Maller, R., Resnick, S., & Shemehsavar, S. (2021b). Splitting the sample at the largest uncensored observation. *arXiv preprint arXiv:2103.01337*.

Maller, R., & Zhou, X. (1996). *Survival analysis with long term survivors*. John Wiley; sons.

Maller, R. A., & Zhou, S. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika*, *79*(4), 731–739.

Maller, R. A., & Zhou, S. (1994). Testing for sufficient follow-up and outliers in survival data. *Journal of the American Statistical Association*, *89*(428), 1499–1506.

Mises, R. v. (1936). The distribution of the largest of n values. *Rev. Math. Interbalcanic Union*, *1*, 141–160.

Müller, P., & Mitra, R. (2013). Bayesian nonparametric inference–why and how. *Bayesian Analysis (Online)*, *8*(2).

Nassar, M., & Elmasry, A. (2012). A study of generalized logistic distributions. *Journal of the Egyptian Mathematical Society*, *20*(2), 126–133.

National Cancer Institute, DCCPS, Surveillance Research Program. (1975-2016). *Surveillance, epidemiology and end results (seer) program research data*. www.seer.cancer.gov

Othus, M., Barlogie, B., LeBlanc, M. L., & Crowley, J. J. (2012). Cure models as a useful statistical tool for analyzing survival. *Clinical Cancer Research*, *18*(14), 3731–6.

Peng, Y. (2003). Fitting semiparametric cure models. *Computational statistics & data analysis*, *41*(3-4), 481–490.

Peng, Y., & Carriere, K. (2002). An empirical comparison of parametric and semiparametric cure models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *44*(8), 1002–1014.

Peng, Y., & Dear, K. B. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, *56*(1), 237–243.

Peng, Y., Dear, K. B., & Denham, J. (1998). A generalized f mixture model for cure rate estimation. *Statistics in Medicine*, *17*(8), 813–830.

Peng, Y., & Taylor, J. M. (2014). Cure models. *Handbook of Survival Analysis*, 113–134.

Peng, Y., & Yu, B. (2021). *Cure models: Methods, applications, and implementation*. Chapman; Hall/CRC.

Pham-Gia, T., & Duong, Q. P. (1989). The generalized beta-and f-distributions in statistical modelling. *Mathematical and Computer Modelling*, *12*(12), 1613–1625.

Prentice, R. L. (1974). A log gamma model and its maximum likelihood estimation. *Biometrika*, *61*(3), 539–544.

Prentice, R. L. (1975). Discrimination among some parametric models. *Biometrika*, *62*(3), 607–614.

Resnick, S. I. (1987). *Extreme values, regular variation and point processes*. Springer.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *The Annals of Statistics*, 303–328.

Sawyer, S. (2003). The greenwood and exponential greenwood confidence intervals in survival analysis. *Applied survival analysis: regression modeling of time to event data*, 1–14.

Schemper, M., & Smith, T. L. (1996). A note on quantifying follow-up in studies of failure time. *Controlled clinical trials*, *17*, 343–346.

Schmidt, P., & Witte, A. D. (1989). Predicting criminal recidivism using 'split population'survival time models. *Journal of Econometrics*, *40*(1), 141–159.

Shen, P.-s. (2000). Testing for sufficient follow-up in survival data. *Statistics & Probability Letters*, *49*(4), 313–322.

Stacy, E. W. (1962). A generalization of the gamma distribution. *The Annals of mathematical statistics*, *33*(3), 1187–1192.

Sundberg, R. (2010). Flat and multimodal likelihoods and model lack of fit in curved exponential families. *Scandinavian Journal of Statistics*, *37*(4), 632–643.

Sy, J. P., & Taylor, J. M. (2000). Estimation in a cox proportional hazards cure model. *Biometrics*, *56*(1), 227–236.

Tai, P., Yu, E., Cserni, G., Vlastos, G., Royce, M., Kunkler, I., & Vinh-Hung, V. (2005). Minimum follow-up time required for the estimation of statistical cure of cancer patients: Verification using data from 42 cancer sites in the seer database. *BMC Cancer*, *5*(1), 48.

Taweab, F., & Ibrahim, N. A. (2014). Cure rate models: A review of recent progress with a study of change-point cure models when cured is partially known. *J. Appl. Sci*, *14*, 609–16.

Taylor, J. M. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics*, *51*(3), 899–907.

Titterington, D., Smith, A., & Makov, U. (1985). *Statistical analysis of finite mixture distributions*. Wiley, New York.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, *72*(1), 20–22.

Tsodikov, A. (1998). A proportional hazards model taking account of long-term survivors. *Biometrics*, *54*(4), 1508–1516.

Tsodikov, A. D., Asselain, B., Fourque, A., Hoang, T., & Yakovlev, A. Y. (1995). Discrete strategies of cancer post-treatment surveillance. Estimation and optimization problems. *Biometrics*, 437–447.

Vu, H., Maller, R., & Zhou, X. (1998). Asymptotic properties of a class of mixture models for failure data: The interior and boundary cases. *Annals of the Institute of Statistical Mathematics*, *50*, 627–653.

Wei, L. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, *11*(14-15), 1871–1879.

Welsh, A. H. (2011). *Aspects of statistical inference* (Vol. 916). John Wiley & Sons.

Weng, J., Yang, D., & Du, G. (2018). Generalized f distribution model with random parameters for estimating property damage cost in maritime accidents. *Maritime Policy & Management*, *45*(8), 963–978.

Williams, J., & Lagakos, S. (1977). Models for censored survival analysis: Constant-sum and variable-sum models. *Biometrika*, *64*(2), 215–224.

Woods, L., Rachet, B., Lambert, P., & Coleman, M. (2009). 'Cure' from breast cancer among two populations of women followed for 23 years after diagnosis. *Annals of Oncology*, *20*(8), 1331–1336.

Xia, F., Ning, J., & Huang, X. (2018). Empirical comparison of the breslow estimator and the kalbfleisch prentice estimator for survival functions. *Journal of Biometrics & Biostatistics*, *9*(2).

Xu, J., & Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Canadian Journal of Statistics*, *42*(1), 1–17.

Yakovlev, A. Y., Asselain, B., Bardou, V., Fourquet, A., Hoang, T., Rochefediere, A., & Tsodikov, A. (1993). A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. *Biometrie et analyse de donnees spatio-temporelles*, *12*, 66–82.

Yamaguchi, K. (1992). Accelerated failure-time regression models with a regression model of surviving fraction: An application to the analysis of "permanent employment" in japan. *Journal of the American Statistical Association*, *87*(418), 284–292.

Yin, G., & Ibrahim, J. G. (2005). A general class of bayesian survival models with zero and nonzero cure fractions. *Biometrics*, *61*(2), 403–412.

Yu, B., Tiwari, R. C., Cronin, K. A., & Feuer, E. J. (2004). Cure fraction estimation from the mixture cure models for grouped survival data. *Statistics in Medicine*, *23*(11), 1733–1747.

Zhang, J., & Peng, Y. (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in Medicine*, *26*(16), 3157–3171.

Zhou, S., & Maller, R. A. (1995). The likelihood ratio test for the presence of immunes in a censored sample. *Statistics: A Journal of Theoretical and Applied Statistics*, *27*(1-2), 181–201.