# Estimation of Survival with a Combination of Prevalent and Incident Cases in the Presence of Length Bias

by

Ewa Makvandi-Nejad

Thesis submitted to the

Faculty of Graduate and Postdoctoral Studies

In partial fulfillment of the requirements

For the M.Sc. degree in

Biostatistics[1]

Department of Mathematics and Statistics

Faculty of Science

University of Ottawa



[1]The program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics.

# Abstract

In studying natural history of a disease, incident studies provide the best quality estimates; in contrast, prevalent studies introduce a sampling bias, which, if the onset time of the disease follows a stationary Poisson process, is called length bias. When both types of data are available, combining the samples under the assumption that failure times in incident and prevalent cohorts come from the same distribution function, could improve the estimation process from a prevalent sample. We verify this assumption using a Smirnov type of test and construct a likelihood function from a combined sample to parametrically estimate the survival through maximum likelihood approach. Finally, we use Accelerated Failure Time models to compare the effect of covariates on survival in incident, prevalent, and combined populations. Properties of the proposed test and the combined estimator are assessed using simulations, and illustrated with data from the Canadian Study of Health and Aging.[2]

[2]The CSHA is detailed in the following 3 papers - Canadian Study of Health and Aging: study methods and prevalence of dementia (1994), The Canadian Study of Health and Aging: risk factors for Alzheimers disease in Canada (1994), The Canadian Study of Health and Aging: patterns of caring for people with dementia in Canada (1994).

## Acknowledgements

*"We learn more by looking for the answer to a question and not finding it than we do from learning the answer itself."*

Lloyd Alexander

Thanks to professor Pierre-Jerôme Bergeron, an amazing supervisor, for making the learning road easier, for his encouragement, guidance, and patience. Also, to all who were and still are in my life for supporting my decisions, helping when in doubts, and accepting even before understanding.

The data used in this thesis were collected as part of the Canadian Study of Health and Aging. The core study was funded by the Seniors Independence Research Program, through the National Health Research and Development Program (NHRDP) of Health Canada (project no. 6606-3954-MC(S)). Additional funding was provided by Pfizer Canada Incorporated through the Medical Research Council/Pharmaceutical Manufacturers Association of Canada Health Activity Program, NHRDP (project no. 6603-1417-302(R)), Bayer Incorporated, and the British Columbia Health Research Foundation (projects no. 38 (93-2) vand no. 34 (96-1)). The study was coordinated through the University of Ottawa and the Division of Aging and Seniors, Health Canada.

# Contents

# List of Tables

# List of Figures

# Chapter 1

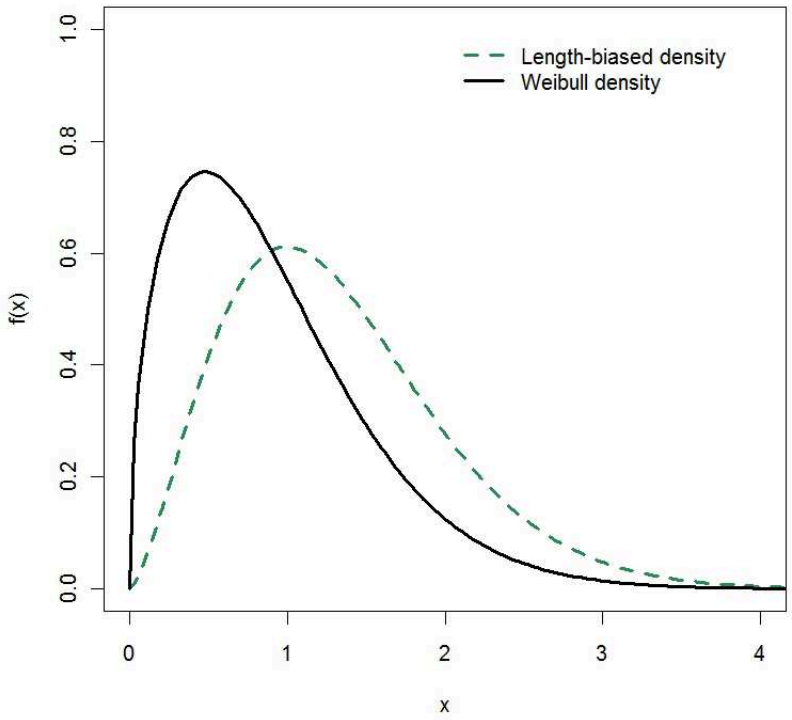
# Introduction

Survival analysis, the study of events involving an element of time, is a relatively new however rapidly developing field of statistics. Its history goes back to the beginning of the development of actuarial science and demography in the 17th century when John Graunt presented the first life table in 1662 [24]. However, it was not until the 1970s when the methodology, theory, and applications were extensively developed mostly due to the work of Mantel [33] and Cox [16]. Over the last decades, the time-to-event approach became an attractive alternative for researchers who were more interested in answering the question of *how long* until an event occurs, rather than *whether* or *not* it occurred in a given time period [21]. Depending on the field of application, survival analysis can be regarded as a lifetime or survival time analysis (clinical and health related studies), reliability analysis (industrial engineering), or event-history analysis (sociology). Regardless of the term used, survival analysis is a collection of statistical procedures for which the time from a defined starting point to the occurrence of a given event is of central interest. *Time* and *event*, the basic terms in survival analysis, can refer to many different situations: in the medical field, they can denote the time interval from the diagnosis of a disease to recovery (event), or time from the beginning to the end of a remission period; in industrial applications, time to failure of a certain mechanical or electronic component; and in sociology, time to marriage (event), or turnover time of new graduate in their first job where the termination of the job is the event of interest [31].

In this thesis, we will look at a special type of survival data, called *length-biased* data, that occur naturally for some sampling plans in biometry, wildlife studies, marketing and health outcome analysis. Length-biased sampling often arises due to economical and time related constraints. For example, in the field of medical research, the ideal setting for describing the course of a disease would be prospective cohort (incident) studies in which healthy subjects are recruited and followed until the onset of a disease, and then followed again until an event of interest occurs. In the case of a rare disease or a disease with a long latent period, this type of study is not always possible from an economical and practical point of view. Instead, less costly, but more efficient sampling strategies such as cross-sectional cohort (prevalent) studies are often implemented. By requiring minimal enrolment and reduced follow-up periods, the prevalent cohort study is more feasible to conduct. Its design, however, introduces systematic bias in the data-generation process, which, if ignored, can lead to incorrect conclusions regarding survival [13]. One form of selection bias that often occurs in prevalent studies is a length-biased sampling in which observations are selected with probability proportional to their *length*. The resulting distribution of lifetime data is then called *length-biased*. In other words, the time intervals actually observed tend to be longer when sampling from length-biased data and ignoring selection effect than those arising from the true distribution of failure times. The probabilities obtained from survival analysis are usually bigger in the presence of length-biased sampling. Consequently, the survival function corresponding to these data is length-biased and differs from the unbiased survival function derived from incident cases [5]. Figure 1.1 illustrates how the length-biased probability density function relates to the probability function derived in the absence of the bias.

To understand the mechanism behind length-biased sampling, let's look at a simple example. Saturday, just before Christmas in Ottawa. People occupy malls in search for gifts. How long can their daily escapades last? Two, three, ten hours? In order to answer this question, we can either follow randomly chosen shoppers from the time they enter the mall to the moment they leave ("survival" shopping time), or we could randomly choose shoppers at certain time and location point, lets say at 2pm on the second floor of the Rideau Centre, and then follow

Figure 1.1: Unbiased and length-biased survival functions




them up until they leave the mall. Of course, in the latter scenario, we are more likely to select those with a longer time in the mall, and miss all ”fast” shoppers or those who came after our ”selection point” was set. While the first case could be regarded as an incident cohort, the second is an example of a prevalent cohort. In incident studies, the ”onset” time (time a person entered the mall) is precisely determined, and all subjects are followed until the event of interest (leaving the mall) occurs. In contrast to incident cohorts, in a prevalent cohort (the second scenario), the data are subjected to selection bias. In other words, for those individuals not included in the study, the initial and terminating events are experienced before the beginning of the study [11]. In the example above, some people enter and exit the mall before we start the experiment. In epidemiological studies, for example in a study of dementia, the subjects who died of rapidly developed dementia before recruitment time are not included (not observable) in the study. Since we only observe individuals who have lived long enough to make it into the

study, the sampling distribution of survival times is said to be left-truncated. In addition, since not all the subjects died before the study ended (or left the mall before we did), the data are also right- censored. While right censoring is a feature of both incident and prevalent samples, left truncation is specific to prevalent data.

Incident studies provide bias-free survival data and allow for seamless inference regarding disease incidence in the target population using standard techniques of survival analysis. In the case of a prevalent study, methods adjusting for the bias need to be employed. Often, in longitudinal studies, information on both, diseased and non-diseased subjects is available as new subjects are recruited as time passes. Therefore, methods that allow an efficient and exhaustive utilization of all the available information in the data are desired. Examples of studies containing both prevalent and incident cohorts include the Canadian Study of Health and Aging (10,000 subjects with dementia, 1991-2001), the National Longitudinal Survey of Children and Youth (data collected on children 0-11 years old every 2 years since 1994), the Canadian Longitudinal Study on Aging (50,000 subjects followed for 20 years), and the British Medical Research Councils cognitive function and ageing study (13,000 subjects followed for 14 years).

Our goal is to parametrically estimate the survival function from large studies that include both incident and prevalent cases. To make use of all the relevant data, there exists only one nonparametric estimator of the survival function to combine these two types of observations, the modified Kaplan-Meier estimator for left-truncated data under unspecified truncation distribution, and it was not developed to that end, thus it does not take advantage of the mathematical properties that distinguish these two types of cases. To justify combining incident and prevalent cases in parametric analysis, we will first assess the distributional equality of the two types of data using existing nonparametric estimators of the survival function: the Kaplan-Meier estimator for unbiased data and the NPMLE estimator developed by Vardi for length-biased data. Once the assumption is verified, we will obtain a parametric MLE based on the combined length-biased and unbiased cases by means of a combined likelihood function. We will then compare precision and efficiency of the estimates from the combined like-

lihood with the MLEs obtained from each type of data, separately. Data on dementia from the Canadian Study of Health and Aging (CSHA) will be used to illustrate our approach and as a benchmark for simulations studies. In Chapter 2, we will introduce basic concepts of survival analysis and maximum likelihood estimation methods for unbiased and length-biased sampling scenarios, followed by a likelihood function for the combined, incident and prevalent, cases. We will then describe methods of estimation of the survival function in prevalent and incident samples. Particularly, we will look more closely at two nonparametric estimators: the Kaplan-Meier estimator and the NPMLE developed by Vardi for the unbiased and length-biased cases, respectively (Chapter 3). In Chapter 4, we will proceed to parametric regression models for survival data, discuss in more detail accelerated failure time models (AFT), and present their adaptation to the two sampling scenarios. We will also derive likelihood functions in the presence of covariates for the three types of lifetime data: incident, prevalent, and combination of incident and prevalent cases. In Chapter 5, we will investigate prerequisites for combining data, the equality of distributions functions, and present a method of comparing the distribution functions in unbiased and length-biased settings using a two-sample Kolmogorov-Smirnoff type of test. Finally, in Chapter 6 and 7, we will validate our approach by performing simulation studies, and apply the proposed methods to the CSHA data in order to estimate the survival function for subjects diagnosed with dementia. Chapter 8 summarizes our estimation efforts, and discusses limitations and possible directions for future research on the discussed subject.

# Chapter 2

# Survival analysis - basic concepts

## 2.1 Survival and hazard functions

In analyzing survival data, two functions that are dependent on time are of particular interest: the survival function and the hazard function. We will focus on the case where time $T$ is a continuous random variable, but survival and hazard functions can also be defined for discrete random variables as well as for a mixture of discrete and continuous random variables.

The survival function $S(t)$ is the complement of the cumulative distribution function $F(t)$. As such, it is a nonincreasing function with a value of 1 at the origin and 0 at infinity. It is defined as the probability of surviving at least to time $t$:

$$S(t) = P(T > t) = \int_t^\infty f(x)dx, \tag{2.1}$$

$T$ here is a nonnegative, continuous random variable and $f(t)$ is the probability density function (pdf) of $t$ [36]. The graph of $S(t)$ against $t$ is called the survival curve.

Since it is often difficult to define the failure pattern by looking at the survival curve alone, the hazard rate is another basic quantity of interest in survival analysis. In contrast to the survival function that focuses on an event not occurring (probability of survival), the hazard function describes the way the risk of failure changes over the time and therefore focuses

on failing. More formally, the hazard function (known also as the hazard rate, the intensity function, the force of mortality, the age-specific failure rate), $h(t)$, is a nonnegative function defined as the conditional probability of dying at time $t$, having survived to that time:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}. \tag{2.2}$$

The hazard function nicely relates to the survival function and vice-versa. From 2.1 we can see that $f(t) = -\frac{d}{dt}S(t)$ and by substituting into 2.2 we obtain the following relations between hazard and survival functions:

$$h(t) = -\frac{d}{dt} \log S(t) \tag{2.3}$$

and

$$S(t) = e^{-\int_0^t h(x)dx} \tag{2.4}$$

where $\int_0^t h(x)dx$ is defined as the cumulative hazard function $H(t)$ [31]. In contrast to the survival function, the shape of the hazard function can assume different forms depending on the risk of failure in a given population. Constant hazard functions, a feature of exponential models, may be used to describe a healthy individuals whose risk of failure is not changing over a certain time period; with increasing hazard functions we can describe patients with a fatal disease whose risk of dying increases over the time; decreasing hazard functions can be used to describe patients recovering from a therapy or surgery. And, when the risk of experiencing an event for individuals is much higher in the first period of a disease but decreases later, the hazard function will be increasing over a certain time period and will decrease after the danger is over [29].

Both survival and hazard functions have their special place in regression models described in details in later chapters. Two widely used regression models in survival analysis, the accelerated failure time models (AFT) and proportional hazard models (PH) use the survival function and the hazard function respectively to model the survival data.

## 2.2 What can go wrong? - censoring and truncation

A key analytical problem in survival analysis is that for some time-to-event data only partial information is available. These times are subjected to censoring and/or truncation. In this section, we will look more closely at the problem of right censoring and left truncation in relation to incident and prevalent data in survival analysis.

According to Cox, there are three requirements for determining survival time: unambiguously defined time origin, appropriately chosen scale for measuring the time's passage, and clearly defined failure [17]. In a perfect world, having the three components for all the subjects in a study would lead to "perfect" survival data and standard regression and estimation methods could be sufficient to obtain "perfect" (i.e. unbiased) estimates of chosen statistics. However, the world of statistics is far from being perfect and defining the time-to-event interval can be challenging. While it is usually relatively easy to define the endpoint of the interval (death, end of marriage, failure of some mechanical component, or end of a study), it is not so straightforward to establish the initial event that triggered the times. In most cases, standard statistical techniques cannot usually be applied because the underlying distribution of time intervals is rarely normal and the data are often *censored* and/or *truncated*.

There are various types of censoring in survival datasets: right, left, and interval censoring. *Right censoring* is the most common type of censoring and occurs when we have partial information about the subject's survival time, but we do not know the survival time exactly (the "complete" survival time is cut off on the right side of the observed time interval). The reason for a subject being right-censored can be one of the following:

- the study ends and the person does not experience the event
- a subject is lost to follow-up during the study period
- a subject withdraws from the study for reason other than the event of interest.

*Left-censoring* occurs when the "complete" survival time has been cut-off at the left side of the observed survival time interval. In other words, the real event's occurrence is not caught

on time and the observed survival time is longer than it really is. The true event time $T$ is not observed but is known to be less than or equal to some time point $T_L$. In the case when all that is known about $T$ is that it is somewhere between two time points, $T_L$ and $T_R$, we say that the observation is *interval-censored*. Interval censoring is just a generalization of left and right censoring [29].

*Truncation* is another common feature of survival studies. As right censoring is more frequent than left censoring, left truncation occurs more often than right truncation. For *left truncation*, only subjects who experienced the initiation of the event (e.g. an onset of a disease) but not yet the event of interest, are included in the study. As opposed to left censoring in which we know that a subject exists but the event is not observed, in left truncation, we may be completely unaware of the subject. For example, people may not be observed until they have reached the age to enter a retirement home. Any deceased subjects in the pre-retirement age group would be unknown. *Right truncation* does not occur as often in practice and is typical for mortality studies where only individuals who already experienced the event (death) are observed [36]. Often in a study, more than one type of censoring or a combination of censoring and truncation occur at the same time. As illustrated in the introduction, cross-sectional (prevalent) studies are subject to both right-censoring and left-truncation. On the other hand, incident studies are naturally free of truncated data as the subjects included in the study did not yet experience onset of the disease. Figures 2.1 and 2.2 illustrate the setting of survival data in incident and prevalent cohorts, respectively. For example, the second observation in figure 2.2 will be not included in the study since both the onset of disease and the event occurred before the study began.

Knowledge of the type of censoring and/or truncation occurring in a dataset is fundamental for inferential purposes. For example, in the process of estimation, the maximum likelihood function depends on the type of data considered.

Figure 2.1: Incident study

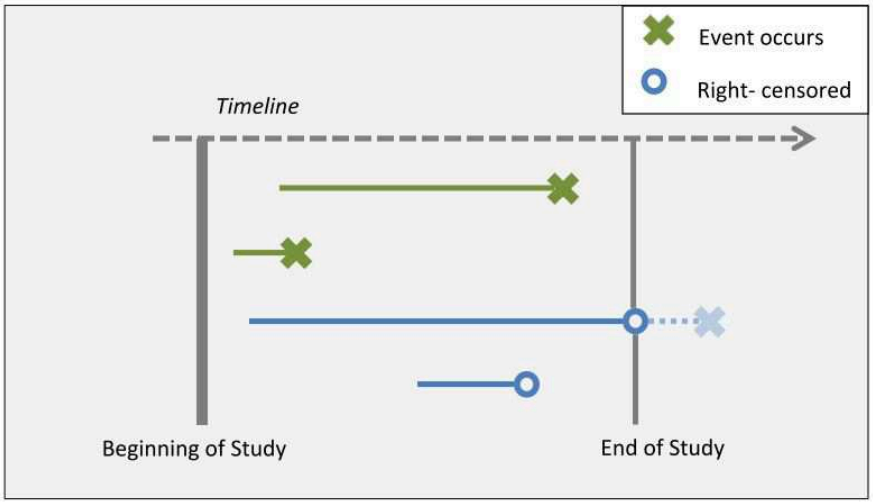
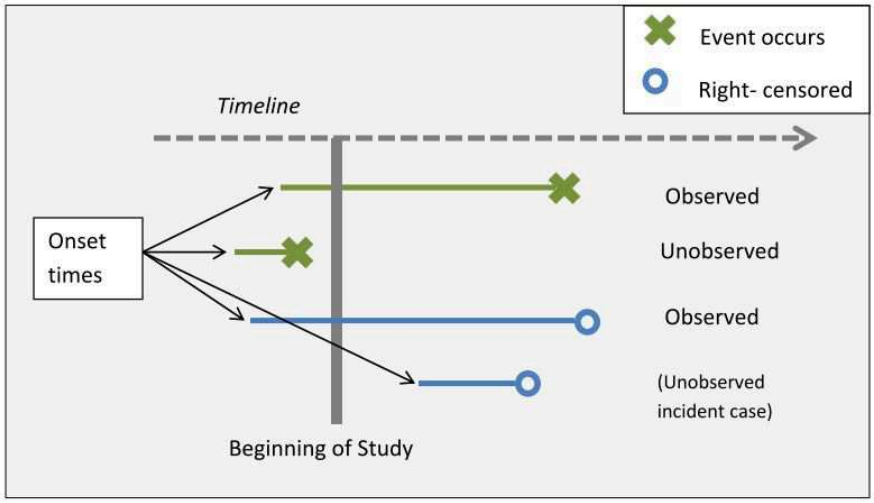



Figure 2.2: Prevalent study



## 2.3 Maximum likelihood estimation procedure for survival data

In this section, we will look at the estimation procedure for survival data by means of a maximum likelihood function. Particularly, we will explain how to obtain maximum likelihood estimates (MLEs) in the presence of length-biased sampling and in the case when both types of data, incident and prevalent, are combined together into one sample.

The likelihood function is a measure of the likelihood of observing a specific set of survival times given $\theta$, where $\boldsymbol{\theta}$ belongs to some parameter space $\Theta$. Here, $\boldsymbol{\theta}$ could be either a real-valued or vector-valued parameter. The likelihood function is the joint *pdf* of the sample regarded as a function of $\boldsymbol{\theta}$ for a given value $x_1, x_2, \ldots, x_n$. The general form of the likelihood function for a random sample of $n$ individuals with the lifetimes $x_1, x_2, \ldots, x_n$, probability density function $f(x)$, and distribution function $F(x)$ is as follows:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i, \boldsymbol{\theta}). \tag{2.5}$$

The maximum likelihood estimator (MLE), denoted by $\hat{\boldsymbol{\theta}}$, is the value of $\boldsymbol{\theta}$ in $\Theta$ that maximizes $L(\boldsymbol{\theta})$ or, equivalently, maximizes the log-likelihood function of $\boldsymbol{\theta}$. This likelihood function can be then maximized to obtain a maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, which is the solution of the simultaneous equations obtained by taking the derivative of $\log L(\boldsymbol{\theta})$ with respect to each $\theta_j$:

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial \theta_j} = 0 \quad j = 1, 2, ...p \tag{2.6}$$

where $p$ is a number of parameters to estimate.

Consequently, estimates of the density function $f(x_i, \boldsymbol{\theta})$ and the distribution function $F(x; \boldsymbol{\theta})$ can be obtained [31]. If the solution to the 2.6 does not have a closed form then numerical methods such as the Newton-Raphson iterative procedure can be used instead [32].

In order to write the likelihood function in the above form, we have to assume that the observations are independent and come from the same distribution, and that the survival and

censored times do not depend on each other. Since the concept of informative and noninformative censoring is crucial for this work (and for utilizing the standard likelihood function), it is important to fully understand it. The estimates for survival functions and their variances rely on independence between censoring times and survival times. If the pattern of censoring is not independent of the survival times, then survival estimates may be biased, and the variance estimates may be inaccurate. In such a case, censoring is regarded as being *informative*. For example, if individuals who are more ill and therefore more likely to die faster tend to leave the study (be censored), then the estimate of the survival function for the population will more likely be too high; if on other hand, individuals who will live longer tend to drop out of the study or are lost to follow-up, the estimates will be too low. So, all we "want" to know for an observation censored at time $x$ under noninformative censoring is that the lifetime exceeds $x$. In addition to noninformative censoring, we also require that the assumption of independence of observations is not violated. That is, the probability of the loss of one subject is not affected by the loss or withdrawal of other subjects, as this may lead to incorrect (biased) estimates of the survival function [23].

Now, we can return to the construction of the likelihood function under different censoring and truncation schemes with respect to the above assumptions. As stated by Moeschberger, to write the likelihood function for survival data we need to consider the information that each of the type of data contributes to the likelihood. The table below summarizes the type of information provided by the data and the corresponding functional form of the contribution [36].

Table 2.1: Contribution of different type of lifetime data to likelihood function

| *Type of lifetime data* | *Description of contribution* | *Function form of contribution* |
|---|---|---|
| exact lifetimes (D) | $P(X = x)$ | $f(x)$ |
| right-censored (R) | $P(X > C_r)$ | $S(C_r)$ |
| left-censored (L) | $P(X < C_l)$ | $1 - S(C_l)$ |
| left-truncated | | $f(x)/S(Y)$ |
| right-truncated | | $f(x)/(1 - S(Y))$ |
| interval-censored (I) | | $S(L_i) - S(R_i)$ |

The components above relevant to a given lifetime data can be then used to construct the likelihood function:

$$L \propto \prod_{i \in D} f(x_i) \prod_{i \in R} S(C_r) \prod_{i \in L} [1 - S(C_l)] \prod_{i \in I} [S(L_i) - S(R_i)]. \tag{2.7}$$

In the case of left truncation, we replace the distribution of exact times $f(x_i)$ and censored times $S(C_i)$ with $f(x_i)/S(Y_i)$ and $S(C_i)/S(Y_i)$, respectively. Note that, in this case, we also assume independence of survival times from truncation times. Otherwise, different techniques of estimation must be applied.

Let's apply the above likelihood building procedure to a real life scenario. Consider a common type of epidemiological survival data collected from an incident study conducted over a specific period of time. For example, patients after a surgery are followed until a remission occurs for the period of five years. There are two possible outcomes for each individual in the study: either we observe remission of the symptoms at some time $X_i$ within the 5 years or the study ends before the event can be observed and those individuals are censored. Of course, other causes of censoring such as loss to follow-up or withdrawal from the study are possible. However, since they do not change the setup for the likelihood function under the assumption of independence between censoring and survival times, we can use the simplified scenario. Each individual in the study has a fixed potential censoring time $C_i$, and $X_i$ is observed only if

$X_i < C_i$. Therefore, $x_i = \min(X_i, C_i)$. Lets $\delta_i$ be an indicator of censoring such that

$$\delta_i = \begin{cases} 1, & \text{if } X_i < C_i \\ 0, & \text{otherwise} \end{cases} \tag{2.8}$$

The joint distribution function $f(x_i, \delta_i)$ is as follows

$$f(x_i, \delta_i) = f(x_i)^{\delta_i} P(X_i > C_i)^{1-\delta_i} = f(x_i)^{\delta_i} S(x_i)^{1-\delta_i} \tag{2.9}$$

[31]. To see this we have to consider two cases: when $\delta = 0$ and when $\delta = 1$:

$$\begin{aligned} P(x_i, \delta_i = 0) &= P(X_i = C_i | \delta_i = 0) P(\delta_i = 0) \\ &= P(\delta_i = 0) = P(X_i > C_i) = S(x_i), \end{aligned} \tag{2.10}$$

and

$$\begin{aligned} P(x_i, \delta_i = 1) &= P(X_i = x_i | \delta_i = 1) P(\delta_i = 1) \\ &= P(X_i = x_i | x_i \leq C_i) P(x_i \leq C_i) \\ &= \frac{f(x_i)}{[1 - S(x_i)]} [1 - S(x_i)] = f(x_i). \end{aligned} \tag{2.11}$$

[36].

The joint likelihood function used to obtain the MLE of $\boldsymbol{\theta}$ will have the following form:

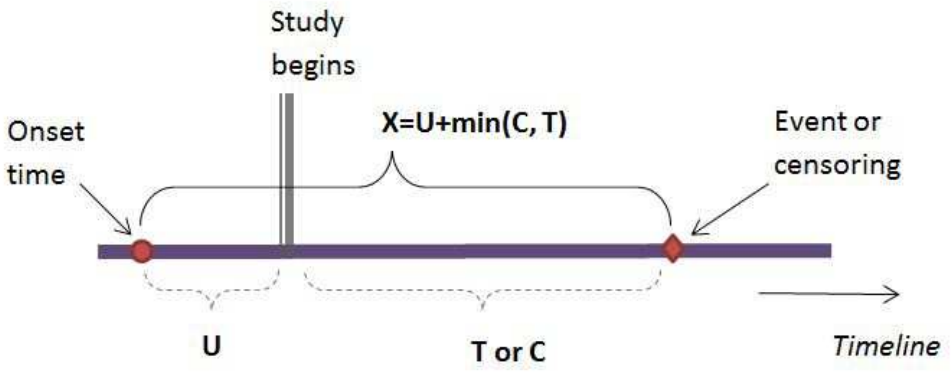
$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i, \delta_i) = \prod_{i=1}^{n} f(x_i)^{\delta_i} S(x_i)^{1-\delta_i} \tag{2.12}$$

The above likelihood function applies when the data come from an incident study. In case of prevalent studies, some of the assumptions listed above are not satisfied and the likelihood function has to be modified.

### 2.3.1 Length-biased distribution and corresponding maximum likelihood function

As mentioned in the previous section, special methods are required in order to obtain unbiased estimates from prevalent cases collected through a cross-sectional survey (i.e. from length-biased sampling). The formal set-up for length-biased data was first derived by [16] and can

Figure 2.3: Failure time distribution in the length-biased sampling




be summarized as follows: for each subject in the study, in addition to a failure time $t_i$ or a censoring time $c_i$, we have also a truncation time $u_i$, the "unobserved" time from the initiation of the event (e.g. from the "onset" of a disease) to the study recruitment time. The "full" time for $i$th subject $x_i$ consists therefore of the observed failure time $t_i$ or censored time $c_i$ depending which one occurs first, and "the unobserved" truncation time $u_i$: $x_i = u_i + \min(t_i, c_i)$ [9]. Figure 2.3 depicts the time composition in the length-biased sampling.

Let X be a random survival time with density function $f_U(x)$ and $U$ be the left truncation time with density function $g(u)$. In order for a subject to be included in the study, $X \geq U$. Under the assumption of independence between survival and truncation times, we can write their joint distribution as

$$f_{X,U}(x, u|X \geq U) = \frac{f_{X,U}(x, u)}{P(X \geq U)} = \frac{g(u)f_U(x)}{P(X \geq U)}, \tag{2.13}$$

where

$$\begin{aligned} P(X \geq U) &= \int_0^\infty P(X \geq u|U = u)g(u)du = \int_0^\infty P(X \geq u)g(u)du \\ &= \int_0^\infty S_X(u)g(u)du \end{aligned} \tag{2.14}$$

In the length-biased sampling scenario, we also assume that the initiation of the event follows a stationary Poisson process and therefore is uniformly distributed on any time interval [8]. This assumption is justified in cases where the phenomenon under the study, for example a

disease, is stable, that is no outbreak of the disease is observed over certain time period. Based on this assumption, the joint distribution of failure and truncation times reduces to $f(x, u|X \geq U) = \frac{f_U(x)}{\mu}$ for $X \geq U$ where $\mu = \int_0^\infty S_X(u)du$ is the mean failure time. The length-biased marginal *pdf* of $x$, $f_{LB}(x)$ is then given by

$$f_{LB}(x|X \geq U) = \int_0^x f(x, u|X \geq U)du = \frac{1}{\mu}\int_0^x f_U(x)du = \frac{xf_U(x)}{\mu} \quad x > 0. \tag{2.15}$$

Note, that when talking about length-biased distribution of $x$ ($f_{LB}(x)$), the unbiased *pdf* and survival function of $x$ will be denoted as $f_U(x)$ and $S_U(x)$ to make a clear distinction between length-biased and unbiased distribution of $x$.

To obtain maximum likelihood estimates for the survival function under the length-biased sampling, one must maximize the appropriate likelihood function.

The derivation of the corresponding approximate length-biased likelihood function will be described in details in the next chapter. Here, we will give its final form. Based on Vardi's approach, the likelihood function for the length-biased data can be approximate by the following:

$$L_{LB}(\boldsymbol{\theta}) = \prod_i^n f_{LB}(x_i)^{\delta_i} S_{LB}(x_i)^{1-\delta_i} \propto \prod_i^n \left(\frac{f_U(x_i)}{\mu(\boldsymbol{\theta})}\right)^{\delta_i} \left(\frac{S_U(x_i)}{\mu(\boldsymbol{\theta})}\right)^{1-\delta_i}, \tag{2.16}$$

where $\boldsymbol{\theta}$ a vector of parameters corresponding to $f_U(x)$, $\mu$ is as before, the mean survival time in the unbiased population, and $\delta_i$ is the indicator of death or censoring for $i$th observation.

### 2.3.2 Maximum likelihood function for combined incident and prevalent data

So far, we have discussed the likelihood function when both prevalent and incident cases were considered separately. Now, we will look at the maximum likelihood function when both types of data are available. The motivation for the combined maximum likelihood is straightforward: why waste data and not to use all of the information it can provide in the process of parameters' estimation? This type of data often arises in longitudinal studies in which we not only have subjects who already experienced the condition of interest, but as the study goes on, new cases

are recruited and followed over time. The Canadian Study of Health and Aging (CHSA) is a good example of such data: the study of dementia consists of both prevalent cases (Phase I) and incident cases (Phase II and III).

Here, we will undertake a parametric approach to derive the likelihood function, i.e. we will assume that the observations and failure times $x_i$ come from the same population with common distribution function $F_U(x)$. This assumption will be later verified using a two-sample Kolmogorov-Smirnoff type test for equality of distributions. Also, since the prevalent and incident data are mutually exclusive i.e. each subject is either a prevalent or an incident case, observations in the combined dataset are assumed to be mutually independent. Let $\xi_i$ be an indicator of whether an observation comes from an incident or prevalent sample, i.e. $\xi_i = 1$ in the incident case and 0 otherwise. Next, let $L_{Inc}$ be the likelihood function for the incident sample given by 2.12, $L_{Prev}$ be the likelihood function for the prevalent sample given by 2.16, and $\boldsymbol{\theta}$ a vector of parameters corresponding to $f(x)$. Then the combined maximum likelihood function can be written as follows:

$$L_{Comb} = L_{Inc}^{\xi_i} \times L_{Prev}^{1-\xi_i}, \tag{2.17}$$

or equivalently,

$$L_{Comb} = \prod_{i=1}^{n} \left[ f_U(x_i)^{\delta_i} S_U(x_i)^{1-\delta_i} \right]^{\xi_i} \left[ \frac{1}{\mu(\boldsymbol{\theta})} f_U(x_i)^{\delta_i} S_U(x_i)^{1-\delta_i} \right]^{1-\xi_i}. \tag{2.18}$$

Simplifying 2.18, we obtain:

$$L_{Comb} = \prod_{i=1}^{n} \left( \frac{1}{\mu(\boldsymbol{\theta})} \right)^{1-\xi_i} f_U(x_i)^{\delta_i} S_U(x_i)^{1-\delta_i} \tag{2.19}$$

Both the unbiased and length-biased likelihood functions, as well as the combined likelihood function, will be used in this thesis to derive the estimates for the survival function under each scenario. However, before we obtain the estimates, let's first look at the methods of estimation of the survival function applicable to our settings.

# Chapter 3

# Nonparametric methods of estimation

Since parametric methods of survival estimation require a strong assumption for the underlying distribution of survival times and can be mathematically complex, survival analysis often relies on nonparametric estimation. In this section, we will describe two nonparametric estimators of the survival function: the standard Kaplan-Meier estimator applicable to incident data, and a more recent estimator developed by Vardi which corrects for the length bias in prevalent data [46].

## 3.1 Kaplan-Meier estimator of survival function

The simplest approach to estimate the survival function is by means of the trivial estimator called the empirical survival function (*esf*). This estimator is based on the empirical cumulative distribution function of the data i.e., a step function that jumps up by $\frac{1}{n}$ at each of the $n$ data points. When lifetime data do not contain censored observations, the *esf* gives reliable estimates; otherwise, it tends to underestimate the survival of subjects, introduce a serious systematic error (bias), cause loss of some information about the sample, and ultimately lower quality of the study [27]. The procedures discussed below do not exclude incomplete observations from the analysis and are therefore more powerful and result in less biased estimates. The nonparametric Kaplan-Meier (KM) estimator (also called product-limit estimator) adjusts the

*esf* to reflect the presence of right-censored observations. The KM estimator was proposed by Kaplan and Meier in 1958 and is still extensively used [28]. As a nonparametric estimator, the Kaplan-Meier method can be used to estimate the survival curve from the observed survival times without the assumption of an underlying probability distribution. The method is based on the basic idea that the probability of surviving $k$ or more periods from entering the study is a product of the $k$ observed survival rates for each period i.e., the cumulative proportion of survival [28]. Let $t_1 < t_2 < ... < t_k$ represent $k$ distinct times at which events occur and let $d_j$ represent the number of events at time $t_j$. In addition, define the number of individuals at risk at time $t_j$, i.e. the number of individuals alive and uncensored before $t_j$, as $n_j$. The KM estimator is then defined as follows:

$$\hat{S}(t) = \prod_{j:t_j<t} \frac{n_j - d_j}{n_j}. \tag{3.1}$$

The Kaplan-Meier estimator corrects for censored observations in the data, but in the case where the data also contain truncated times, the estimate obtained based on the KM method will most likely overestimate the survival of subjects by ignoring those whose survival was too short to be captured by the study [31]. An extension of the KM estimator, $\hat{S}(t)$, a MLE estimator of the cumulative hazard function, $\hat{H}(t) = \sum_{j:t_j<t} \frac{d_j}{n_j}$, was proposed by Nelson and Aalen in 1969 and 1972, respectively [37] [1]. Both estimators possess desirable large-sample properties such as consistency and asymptotic normality under the assumption of noninformative censoring. The asymptotic normality or nonparametric bootstrap techniques can be used to construct variance and confidence intervals for the lifetime distribution. For details on the derivation of asymptotic properties of the above estimators one can refer to [2] and [22]. The KM estimator although useful in estimating survival in many lifetime scenarios such as incident cohort sampling, has to be carefully applied to data which feature both right censoring and left truncation. The standard approach in this case involves conditioning on surviving beyond time $t$ given survival to time $t_L$, the entry point of the study [36]. This approach however does not avoid the problem of the relatively large variance of the KM estimator for small $t$ when data are left-truncated. In 1991, Lai and Ying proposed a conditional modified product-limit

estimator which corrects the conditional KM estimator by ignoring deaths when the size of the risk set is very small [30]. There are some methodological issues with both modified KM estimators: KM's large variability for small and large $t$ where the risk sets are small may affect the estimation of the entire curve [36]; the modified KM puts no weight on the censored observations from prevalent data and its risk set may drop to 0 even though some subject are still alive. In addition, when an assumption on the distribution of truncation times can be made, the conditional approach is less efficient.

In 1986, another conditional approach to estimate right-censored and left-truncated data was proposed by Wang. His idea was to use a conditional nonparametric likelihood estimator (NPMLE), obtained by maximization of the conditional likelihood that would have been obtained if each subjects censoring time been available even if failure occurred before censoring [43]. However, in the situation where an assumption about the truncation times can be made, a conditional approach was proved to be less efficient in contrast to the unconditional approach which proceeds under partial knowledge on the truncation distribution pioneered by Vardi [45]. Conditional and unconditional approaches for estimating the survival function under left truncation are well-discussed in [5].

## 3.2 Length-biased sampling - Vardi approach to estimation of survival function

In 1989, Vardi derived an unconditional, nonparametric maximum likelihood estimator (NPMLE) for the length-biased distribution from multiplicatively right-censored length-biased data [46]. Since in this thesis Vardi's algorithm is extensively used, we will discuss it in detail in this section.

To be able to apply Vardi's approach to our setting we first must make a few assumptions. First of all, we have to assume that the disease under study is stable, i.e., incidence of the disease follows a stationary Poisson process. As a result, the occurrences of disease (truncation times) are uniformly distributed on any time interval. Under this scenario, the survival

times have a length-biased distribution with longer times having a higher probability of being selected [26]. We further assume that the distribution of time from the onset of a disease to the event of interest (death or censoring) is independent of the calendar time of the onset times. Notice also, that in this setting censoring is no longer noninformative but carries some information regarding the survival time of a subject i.e., it depends on the outcome the subject would have had in the absence of censoring.

Let $B$ denote the beginning of the study; $O$, the calendar time of disease's onset; $X$, the time from the onset to failure; and $C$, the time from the onset to censoring. Also, let $U = B - O$ be the truncation time from onset to the sampling time, and $V_x = X - U$ and $V_c = C - U$ be the residual times from the beginning of the study to the failure or censoring, whichever occurs first. In order for a subject to be observed in the study, $X \geq U > 0$ for each individual. Let $f(x)$ and $g(u)$ be the marginal density functions of $X$ and $U$. Then, under the specifications above, the joint $pdf$ of $X$ and $U$, can be written as

$$f_{X,U}(x,u) = \frac{xf(x)}{\mu}\frac{1}{x} = \frac{f(x)}{\mu}, \qquad X > U > 0. \tag{3.2}$$

where $\mu = \int_0^\infty wf(w)dw$ is the mean of $f(x)$ [26]. As shown in the previous chapter, the marginal distribution of failure times $X$ under this scenario has a length-biased distribution function 2.15.

Now, we will show how Vardi's approach can be used to maximize the likelihood function arisen from the above set up.

According to Vardi, estimation of the length-biased distribution is equivalent to estimation of the common underlying lifetime distribution of Vardi's multiplicative censoring model. Although the probability specification in the two cases may differ, and therefore the limiting behaviors of the estimates should be derived separately, the derivation of the likelihood function is the same. Following Vardi, let $X_1, \ldots, X_m$ and $Z_1, \ldots, Z_n$ be identically distributed random variables from a distribution $G$, and $U_1, \ldots, U_n$ be identically distributed uniform $(0, 1)$ random variables. Let $X_i$ denote the complete observation from $G$ and let $Y_i \equiv Z_i U_i$ be the incomplete (censored), scaled down, random variables arisen from multiplying $Z$ and $U$. The process of obtaining $Y$'s is referred to as a "multiplicative censoring" [46]. Since $Y_i \equiv Z_i U_i$,

conditioning on the value of $Z_i = z_i$, $Y$ has uniform $(0, Z)$ distribution and its density function can be written as

$$f_Y(y) = P(Y|Z) = \int_{z \geq y} P(Y|Z = z)g(z)dz = \int_{z \geq y} \frac{1}{z} g(z)dz. \tag{3.3}$$

In order to find the distribution function $G$ we have to maximize the following likelihood function for the full data i.e., data that include both complete $X$s and incomplete $Y$s:

$$L(G) = \prod_{i}^{m} g(x_i) \prod_{i}^{n} \int_{z \geq y_i} \frac{1}{z} g(z)dz. \tag{3.4}$$

From the discrete data point of view we can write the above likelihood in the following form:

$$L(p) = \prod_{j}^{h} p_j^{\xi_j} \left( \Sigma_{k=j}^{h} \frac{1}{t_k} p_k \right)^{\zeta_j}. \tag{3.5}$$

Here $t_1, ..., t_h$ are distinct values of $x_1, \ldots, x_m, y_1, \ldots, y_n$ in increasing order with $h < m + n$ due to possibility of ties in the data, and $\xi_i$ and $\zeta_i$ denote multiplicity of $x_i$ and $y_i$, respectively. Note also that $p_j \geq 0$, $\sum_{j=1}^{h} p_j = 1$, and $p_j \equiv p(t_j) = g(t_j)$.

To maximize the likelihood function, Vardi uses an iterative method that consists of an expectation step ($E$-step) followed by a maximization step ($M$-step), called the $EM$ algorithm. The $EM$ algorithm is a simple method often applied to situations when data include incomplete observations [19]. The general properties of $EM$ algorithm can be summarized as follows:

- the likelihood function increases with each iteration and is guaranteed to converge monotonically to the NPMLE of a distribution function $F$
- the algorithm converges to a unique point $\hat{p}$ that maximizes the above likelihood.

Once the estimates of $p(t_i)$ are obtained, Vardi's NPMLE of survival function $S$ can be written as

$$\hat{S}(t) = \sum \hat{p}(t_i) I(t_i \leq t). \tag{3.6}$$

Now, in order to explain how Vardi's approach can be applied to our scenario, we will look at the setting for prevalent data in terms of a stationary renewal process. Let $A$ denote the

backward recurrence time (truncation time in our setting) and $R$ the forward recurrent time (residual time) [46]. If we write the full survival times as $x_i = a_i + r_i$ for $i = 1, \ldots, m$, and the incomplete (censored) times as $y_i = a_{m+i} + L$ for $i = 1, \ldots, n$ and some time $L$ such that $R_i \leq L$ for "full" observations and $R_i > L$ for incomplete observations, the likelihood function for this data can be written then as

$$\prod_i^m \frac{f(a_i + r_i)}{\mu} \prod_i^n \frac{S(a_{m+i} + L)}{\mu} = \prod_i^m \frac{f(x_i)}{\mu} \prod_i^n \frac{S(y_i)}{\mu}. \tag{3.7}$$

Furthermore, by making the following substitution in 3.4: $g(x) = \frac{xf(x)}{\mu}$ and $\frac{S(y)}{\mu} = \int_{z \geq y_i} \frac{1}{z} g(z)dz$, and noticing that by multiplying $g(x)$ by $x$, the maximum likelihood estimate does not change since the $x$s are treated as constants here, the equivalence of the above likelihood to the likelihood function for Vardi's multiplicative censoring model becomes clear. Under Vardi's approach, the contribution of the distribution function of failure times in length-biased sampling is the same as the contribution of the distribution function of Vardi's full $Z$ variables. Also, the uniformly truncated variables have the distribution function equivalent to Vardi's $ZU$ variables. In this sense, Vardi's approach can be applied to length-biased sampling.

It should be noticed that recently, an alternative estimator for length-biased sampling was proposed by Huang and Qin. This estimator has a small loss of efficiency and eliminates some difficulties associated with the NPMLE estimator proposed by Vardi, such as the lack of a closed-form expression for the asymptotic variance [26].

# Chapter 4

# Regression models for survival data and MLEs in the presence of covariates

## 4.1 Regression models

There are two key concepts that shaped modern survival analysis: Kaplan-Meier's product-limit estimator of the survival function (1958), and Cox's proportional hazard model for modeling the relationship between survival and covariates (1972) [38]. Since their occurrence, nonparametric models, semiparametric models, and methods based on the theory of counting processes and martingales dominated survival literature. In this chapter, we will refocus our attention on parametric models, which, if feasible, have advantages over nonparametric or semiparametric models. Nonparametric methods are quite easy to understand and apply, but they are less efficient than parametric methods when survival times follow a specific distribution, and more efficient when a suitable distribution cannot be determined. Parametric models, if found to be adequate, lead to more precise estimation of the survival probabilities, can be applied to studies with smaller sample sizes, and can help to understand a complex problem by simplifying and synthesizing versus just describing it. For example, in some setting such as clinical trials of cancer therapy, nonparametric methods pose a serious limitation. They are sensitive to differences in the survival between treatment groups, but give no insight into the

mechanisms by which therapy enhances survival [25]. Furthermore, as noted by Bergeron, AFT models come as a more natural choice when length-biased sampling is present [8]. The marginal distribution of covariates in AFT models depends only on regression parameters, and unlike in the Cox proportional hazard model in which the proportionality is lost in the length-biased sample, AFT models can be extended from unbiased to the length-biased population. It is often left to the researcher to decide which of the methods would best describe the data. When one is willing (and able) to assume a parametric form for the distribution of survival times, there exist few common distributions useful in describing the survival data. Despite the fact that parametric models provide a full description of the data they modeled, they are used less frequently mostly due to their lack of flexibility in choosing the appropriate distribution for given data.

In this chapter, we will focus on one form of the parametric models: accelerated failure time models (AFT), discuss in more details the most widely used distribution functions in this model: Weibull and log-normal, and explain conditional and unconditional approaches to estimating parameters in AFT models in the presence of covariates. At the end of this section, we will also mention the proportional hazards model (PH) mostly because of its semiparametric nature and popularity. Other, less frequently used, models such as proportional odds model [34], semiparametric AFT models [6], and accelerated hazard models [14] also exist in the literature but will not be discussed here.

### 4.1.1 Some important distributions

Since our interest lies in parametric models, let's first look at the families of distributions often used to describe survival times: exponential, Weibull, gamma, and log-normal. The first three distributions come from the generalized gamma family and can be described together. The exponential distribution has more of a historical value. It was introduced and widely used in the industrial field, but due to its constant hazard rate (no aging) and memoryless property, its utility in studies when human subjects and history of their disease are of interest is limited. Weibull, log-normal, and log-logistic distributions, on the other hand, do not assume a constant

hazard rate and therefore play a key role in the parametric approach to estimating survival of subjects [32]. We will focus our attention on Weibull and log-normal distributions as they proved to be the best parametric choices for the data used in this thesis: the CSHA study of dementia [48]. The unbiased (incident data) and length-biased (prevalent data) forms of the distributions are of main interest here.

**Generalized gamma distribution**

Let $X$ be a r.v. from the generalized gamma distribution, $GG(k, \lambda, \alpha)$, with the unbiased probability density function given by

$$f_U(x) = \frac{\lambda\alpha(\lambda x)^{\alpha k-1}\exp(-\lambda x)^{\alpha}}{\Gamma(k)}, \tag{4.1}$$

where $\lambda, \alpha, k > 0$ are the parameters of $GG$ and $\Gamma$ is the gamma function. We will use the subscript $U$ to emphasize the unbiased form of a distribution. When $\alpha = 1$, the corresponding function will be gamma $(k, \lambda)$; when $k = 1$, Weibull $(\lambda, \alpha)$; and when $k = \alpha = 1$, exponential $(\lambda)$ [15]. The probability density function $f_U(x)$, the survival functions $S_U(x)$, and the hazard functions $h_U(x)$ for the exponential, Weibull, and gamma distributions are given in Table 4.1.

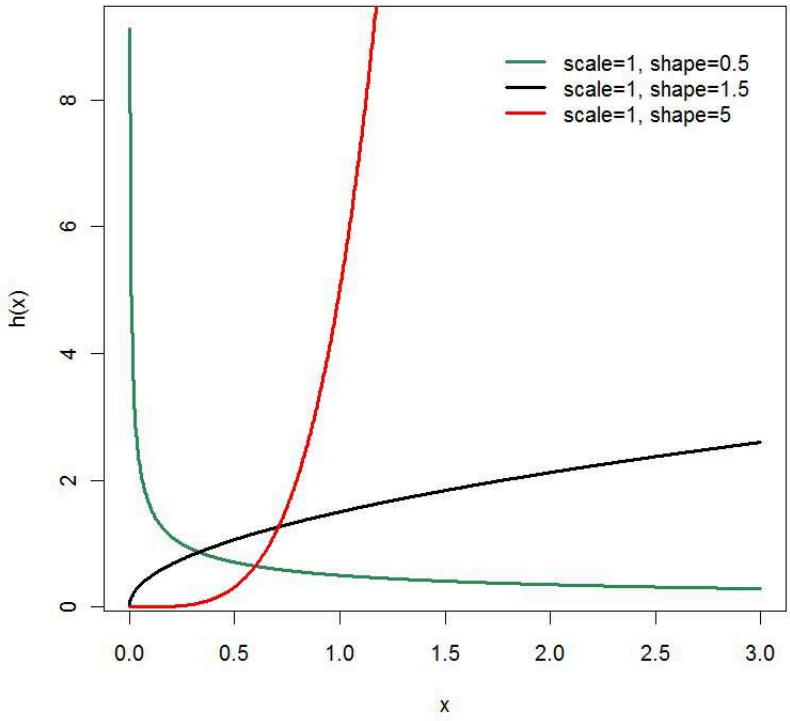
Table 4.1: Hazard rates, survival functions, and probability density functions for exponential, Weibull, and gamma distributions

| *Distribution* | *Hazard rate* | *Survival function* | *Probability Density Function* |
|---|---|---|---|
| | $h_U(x)$ | $S_U(x)$ | $f_U(x)$ |
| Exponential | $\lambda$ | $e^{-\lambda x}$ | $\lambda e^{-\lambda x}$ |
| $\lambda > 0, x \geq 0$ | | | |
| Weibull | $\alpha\lambda x^{\alpha-1}$ | $e^{-\lambda x^{\alpha}}$ | $\alpha\lambda x^{\alpha-1}e^{-\lambda x^{\alpha}}$ |
| $\alpha, \lambda > 0, x \geq 0$ | | | |
| Gamma | $\frac{f(x)}{S(x)}$ | $1 - I(\lambda x, \beta)$ | $\frac{\lambda^{\beta}x^{\beta-1}e^{-\lambda x}}{\Gamma(\beta)}$ |
| $\beta, \lambda > 0, x \geq 0$ | | | |

The shape of the density and hazard functions of the Weibull distribution depends solely on

the shape parameter $\alpha$. Therefore, the Weibull hazard function increases for $\alpha > 1$, decreases for $\alpha < 1$, and becomes constant if $\alpha = 1$. The scale parameter, $\lambda$, affects only the horizontal axis (time) (Figure 4.1. This property explains the robustness of the Weibull distribution in medical application where, for example, the risk of dying for patients with leukemia not responding to treatment increases with time (increasing Weibull hazard) or when the risk of dying after surgery decreases with time (decreasing Weibull hazard) [36].

Figure 4.1: Weibull hazard functions for different values of the shape parameter $\alpha$




Recall that if $Y$ comes from a length-biased distribution, the relation of the probability distribution function of $Y$ to the corresponding unbiased distribution function of $X$ is given by

$$F_y(x) = \int_0^x t f_U(t) dt/\mu \quad t \geq 0, \tag{4.2}$$

and the pdf of Y is

$$f_y(x) = x f_U(x)/\mu \quad t \geq 0. \tag{4.3}$$

The unbiased mean $\mu$ for Weibull r.v. (necessary for deriving the length-biased distribution) is given by

$$\mu = \frac{\Gamma(1+\frac{1}{\alpha})}{\lambda}. \tag{4.4}$$

Substituting $f_U(x)$ (see Table 4.1) and 4.4 to 4.3, we can write the probability density function of the length-biased Weibull r.v as

$$\begin{aligned} f_Y(x) &= \frac{1}{\Gamma(1+\frac{1}{\alpha})}\lambda\alpha(\lambda x)^{\alpha-1}e^{-(\lambda x)^{\alpha}} \\ &= \lambda\alpha(\lambda x)^{\alpha(1+\frac{1}{\alpha}-1)}e^{-(\lambda x)^{\alpha}}\frac{1}{\Gamma(1+\frac{1}{\alpha})}, \end{aligned} \tag{4.5}$$

and recognize it as the *pdf* of a $GG(1+1/\alpha, \lambda, \alpha)$ [15].

Similarly, one can derive the length-biased forms of the exponential and gamma distributions. It is interesting to note that the length-biased distributions of exponential and gamma r.v.s are gamma $(2, \lambda)$ and gamma $(k+1, \lambda)$, respectively [15].

**Log-normal distribution**

The log-normal distribution was studied extensively since 1879 and its application in the medical research traces back to 1940s. Due to its shape and relation to the normal distribution, it became popular for describing the distribution of survival times of many diseases (i.e. Hodgkins's disease or chronic leukemia), where potential of experiencing the event (death) increases in early stages of the disease and decreases later [32].

The log-normal distribution can be characterized as follows: if we let $X$ be a log-normal r.v., then $\log(X)$ will be normally distributed with mean $\mu$ and variance $\sigma^2$. The unbiased density, survival, and hazard functions of the log-normal distribution are summarized in the Table 4.2.

The mean of the unbiased log-normal distribution is given by

$$\mu = e^{\mu+0.5\sigma^2}. \tag{4.6}$$

Similarly to the Weibull distribution, we can derive the length-biased distribution for log-normally distributed lifetimes using 4.3, 4.6, and the unbiased pdf of log-normal:

Table 4.2: Hazard rates, survival function, and probability density function for log-normal distribution

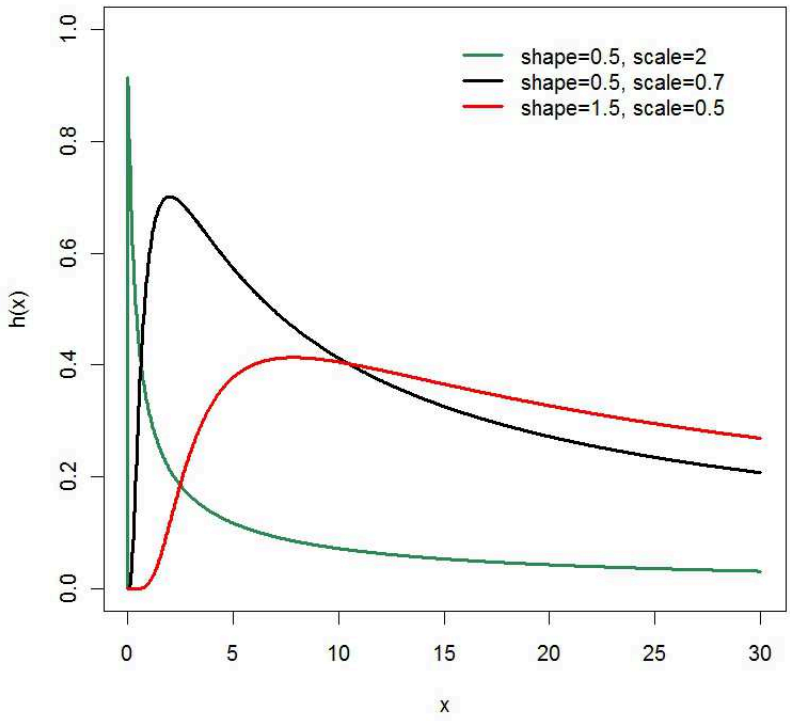
| *Distribution* | *Hazard rate* | *Survival function* | *Probability Density Function* |
|---|---|---|---|
| | $h_u(x)$ | $S_U(x)$ | $f_U(x)$ |
| Log-normal | $\frac{f(x)}{S(x)}$ | $1-\Phi\left[\frac{lnx-\mu}{\sigma}\right]$ | $\frac{1}{x\sqrt{2\pi}\sigma}e^{-0.5(\frac{lnx-\mu}{\sigma})^2}$ |
| $\sigma>0, x\geq 0$ | | | |

$$f_{LB}(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-0.5(\frac{lnx-\mu}{\sigma})^2-\mu-0.5\sigma^2}. \tag{4.7}$$

It it interesting (and useful for the simulations of length-biased log-normal r.v.$s$) to note that if X is log-normal r.v. and we let $Y=\log(X)$, then $Y$ will be normally distributed with mean $\mu_Y=\mu+\sigma^2$ and variance $\sigma_Y^2=\sigma^2$.

The hazard function of log-normal distribution has value 0 at time $x=0$, increases to some maximum, and then decreases, approaching 0 as $x$ become large (Figure 4.2).

Figure 4.2: Log-normal hazard functions for different values of the shape and scale parameters, $\mu$ and $\sigma$




**Log-logistic distribution**

The log-logistic distribution is not utilized in this thesis; however, due to its importance in survival analysis we will mention it here.

A lifetime $X$ has a log-logistic distribution if $Y = \log(X)$ is logistically distributed with the shape parameter $\alpha$ and the scale parameter $\beta > 0$. The log-logistic is a symmetric distribution with slightly heavier tails than the standard normal distribution. The popularity of the log-logistic model in survival analysis is due to its simple expressions for the survival and hazard functions. Its hazard function is monotone decreasing from infinity if $\alpha < 1(\beta > 1)$, monotone decreasing from $\lambda$ if $\alpha = 1(\beta = 1)$, and similar to the log-normal if $\alpha > 1(\beta > 1)$.

All of the distributions mentioned above (except for the gamma distribution) share the

following property: the distribution of the log-lifetime, $\log(X)$, is a member of the location and scale family of distributions and can be expressed in the form

$$Y = \log(X) = \mu + \sigma W, \tag{4.8}$$

where $W$ has an extreme minimum value, normal, or logistic distribution for Weibull, log-normal, and log-logistic distributions, respectively. As we will see in the next section, this property can be easily extended to survival regression models with covariates.

### 4.1.2 AFT models

Accelerated failure time (AFT) and proportional hazard models are the most popular regression models used for modeling survival data. When the assumption of proportionality in proportional hazard models cannot be satisfied and/or the effect of covariates may be better described by an accelerated life, AFT models are the better choice. Parametric AFT models assume that survival times come from a given theoretical distribution and is explicitly related to the covariates [32]. The way the relationship between $X$ and covariates is described differentiates between the types of models used. The AFT model belongs to the class of log-linear models which assumes a linear relationship between $\log(X)$ and the covariates:

$$Y = \log(X) = -\boldsymbol{\beta}'\mathbf{z} + \sigma W, \tag{4.9}$$

where $\sigma$ is an unknown scale parameter and $W$, the error term, is a random variable with known density function. If $W$ assumes an extreme value, normal, or logistic distribution then $X$ has Weibull, log-normal, or log-logistic distribution, respectively.

Let $X$ denote the failure time, $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$ the vector of regression parameters, and $\mathbf{z}' = (z_1, z_2, \ldots, z_p)$ the vector of covariates. Covariates here can be quantities such as age, gender, or whether or not an individual is under treatment. Their purpose in the model is to explain some of the variability in survival time between individuals. The error term can be viewed as a standard or a reference distribution that applies when $Z = 0$. In this case, the distribution of $X$ reduces to $X_0 = e^{\sigma W}$.

In AFT models, the covariates change (accelerate or decelerate) the time-scale of the event under study, i.e. they speed up or slow down the time scale [31]. To illustrate this assumption, let's compare the lifespan of humans and dogs. It is believed that on average the lifetime of human is seven times longer than that of a dog. So, in terms of the survival function, $S_{Dog}(x) = S_{Human}(7x)$. Therefore, two years of human life are equivalent to 14 years of dog's life. In general, for any two subpopulations, the above relation can be expressed as $S_2(x) = S_1(\gamma x)$ where $\gamma$ is a constant called *the acceleration factor*, and $x$ is a time variable. If we parameterize $\gamma$ as $e^{\boldsymbol{\beta}z}$, then the multiplicative effect of covariates on survival time can be written as shown in [29]:

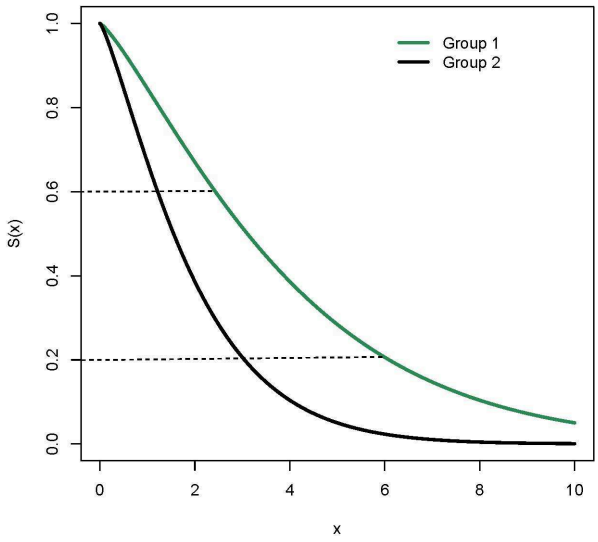
$$X = X_0 e^{\boldsymbol{\beta}'\mathbf{z}}, \tag{4.10}$$

and the corresponding survival function (the probability that the subject with covariate value $z$ will be alive at time $x$) is then given by:

$$S_U(x|\mathbf{z};\boldsymbol{\beta}) = P(X > x|\mathbf{z};\boldsymbol{\beta}) = P(X_0 e^{\boldsymbol{\beta}'\mathbf{z}} > x) = P(X_0 > xe^{-\boldsymbol{\beta}'\mathbf{z}}) = S_0(xe^{-\boldsymbol{\beta}'\mathbf{z}}). \tag{4.11}$$

$S_0$, here, is the baseline survival function when all of the covariates are equal to zero. It is easy to see now that the covariates change the original time scale by a factor of $e^{-\boldsymbol{\beta}'\mathbf{z}}$ and that the time is either accelerated or degraded, depending on the sign of $\boldsymbol{\beta}'\mathbf{z}$. Also, for any fixed value of $S(x)$ (any quantile of $S(x)$ for this matter), the ratio of survival times (the acceleration factor) is constant, i.e., the ratio of distances between two survival curves measured on the horizontal axis (time) is the same at any time $x$. As shown in Figure 4.3, the distance of the horizontal line from $S(x)$ axis to the survival curve for Group 1 is twice as long as the same distance for Group 2 (time ratio (TR) =2).

Figure 4.3: Relationship between survival curves in the AFT models for a binary covariate $\beta$




Based on 4.11, the probability density function for AFT models with covariates can also be written in terms of the baseline density function [32]:

$$f_U(x|\mathbf{z};\boldsymbol{\beta}) = f_0(xe^{-\boldsymbol{\beta}'\mathbf{z}})e^{-\boldsymbol{\beta}'\mathbf{z}}. \tag{4.12}$$

Let's now consider the length-biased case. The probability density function of $x$ given a vector of covariates $\mathbf{z}$ in the presence of length bias will assume the following form:

$$f_{LB}(x|\mathbf{z};\boldsymbol{\beta}) = \frac{xf_U(x|\mathbf{z};\boldsymbol{\beta})}{\mu(\mathbf{z};\boldsymbol{\beta})}, \tag{4.13}$$

where $\mu(\mathbf{z};\boldsymbol{\beta})$, the mean survival time for a given value of covariate $Z$, is calculated as follows:

$$\mu(\mathbf{z};\boldsymbol{\beta}) = \int_0^\infty S_U(x|\mathbf{z};\boldsymbol{\beta})dx = \int_0^\infty S_0(xe^{-\boldsymbol{\beta}'\mathbf{z}})dt = e^{\boldsymbol{\beta}'\mathbf{z}}\int_0^\infty S_0(w)dw = e^{\boldsymbol{\beta}'\mathbf{z}}\mu(0) \tag{4.14}$$

The length-biased density function with covariates can be alternatively written as:

$$\begin{aligned} f_{LB}(x|\mathbf{z};\boldsymbol{\beta}) = \frac{xe^{-\boldsymbol{\beta}'\mathbf{z}}f_{U,0}(xe^{-\boldsymbol{\beta}'\mathbf{z}})}{e^{\boldsymbol{\beta}'\mathbf{z}}\mu(0)} &= \left(\frac{xe^{-\boldsymbol{\beta}\mathbf{z}}f_{U,0}(xe^{-\boldsymbol{\beta}'\mathbf{z}})}{\mu(0)}\right)e^{-\boldsymbol{\beta}'\mathbf{z}} \\ &= f_{LB,0}(xe^{-\boldsymbol{\beta}'\mathbf{z}})e^{-\boldsymbol{\beta}'\mathbf{z}}. \end{aligned} \tag{4.15}$$

Consequently, the length-biased survival function has the following form:

$$S_{LB}(x|\mathbf{z};\boldsymbol{\beta}) = S_{LB,0}(xe^{-\boldsymbol{\beta}'\mathbf{z}}). \tag{4.16}$$

When there are no covariates, the density function assumes the usual length-biased form, $f_{LB,0}(x) = xf_0(x)/\mu(0)$ [8].

Based on 4.16, we can easily see that in the length-biased scenario, the survival function satisfies the assumptions of the AFT models, i.e. as in the unbiased case, the covariates modify the survival time by a factor $e^{-\boldsymbol{\beta}'\mathbf{z}}$. As we will see later, this relationship is not true for the proportional hazard models.

Table 4.3 contains the density and survival functions in the length-biased setting for the two distribution used in this thesis, Weibull and log-normal.

Table 4.3: Length-biased probability density and survival functions for Weibull and log-normal distributions

| *Distribution* | *Survival function* $S_{LB}(x)$ | *Length-biased pdf* $f_{LB}(x)$ |
|---|---|---|
| Weibull $\alpha, \lambda > 0, x \geq 0$ | $e^{-(\lambda x e^{-\boldsymbol{\beta}'\mathbf{z}})^{\alpha}}$ | $\alpha\lambda^{\alpha}x^{\alpha-1}e^{-\boldsymbol{\beta}'\mathbf{z}\alpha}e^{-(\lambda x e^{-(\lambda x e^{-\boldsymbol{\beta}'\mathbf{z}})^{\alpha}})}$ |
| Log-normal $\sigma > 0, x \geq 0$ | $1 - \Phi\left(\frac{\log x - \boldsymbol{\beta}'\mathbf{z} - \mu}{\sigma}\right)$ | $\frac{1}{x\sqrt{2\pi}\sigma}e^{-0.5(\frac{\log x - \boldsymbol{\beta}'\mathbf{z} - \mu}{\sigma})^2}$ |

### 4.1.3 Proportional Hazard model

Another approach to modeling effect of covariates on survival time is to model the hazard rate as a function of the covariates (i.e. $h(x|\mathbf{z})$). In this model, the response (dependent) variable is 'the hazard', i.e. the probability of dying (or experiencing the event in question) given that individual has survived up to a given point in time. The conditional hazard rate for an individual with covariate vector $\mathbf{z}$ can be expressed as a product of the baseline hazard $h_0(x)$ (when all the covariates are equal 0), and some function of $\mathbf{z}$, say $g(\boldsymbol{\beta}'\mathbf{z})$. The hazard function describes how the hazard (risk) changes over time at baseline levels of covariates, and the part containing parameters' effect describes how the hazard varies in response to explanatory covariates:

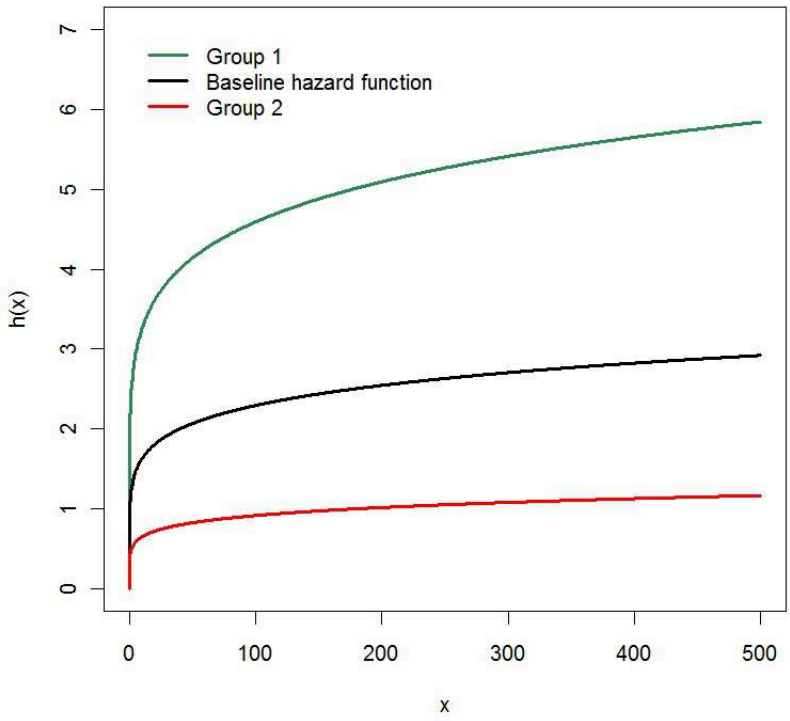
$$h(x|\mathbf{z};\boldsymbol{\beta}) = h_0(x)g(\boldsymbol{\beta}'\mathbf{z}). \tag{4.17}$$

The main assumption of proportional hazard models when all of the covariates are fixed at time $x$, is the proportionality of the hazard rates of two individuals with distinct covariate vectors. Suppose we have two individuals with covariate vectors $\mathbf{z_1}$ and $\mathbf{z_2}$. The ratio of their hazards is

$$\frac{h(x|\mathbf{z_1})}{h(x|\mathbf{z_2})} = \frac{h_0(x)g(\boldsymbol{\beta}'\mathbf{z_1})}{h_0(x)g(\boldsymbol{\beta}'\mathbf{z_2})} = \frac{g(\boldsymbol{\beta}'\mathbf{z_1})}{g(\boldsymbol{\beta}'\mathbf{z_2})}, \tag{4.18}$$

which does not depend on $x$. This quantity can be seen as the relative risk of an individual with covariate vector $\mathbf{z_1}$ experiencing the event of interest compared to an individual with covariate vector $\mathbf{z_2}$ [36]. In other words, covariates are multiplicatively related to the hazard. For example, an exposure to a risk factor A may double a subject's hazard at any given time, while the baseline hazard may vary. The proportionality of hazard rates for different value of a covariate are shown in Figure 4.4. It is necessary to check that the proportional hazards assumption holds before using this type of model. If so, the effect of the parameters can be estimated without considering the form of the hazard function.

Figure 4.4: Relationship between hazard functions in the PH models for different values of covariate $\beta$




Any non-negative function may be used for $g(\boldsymbol{\beta}'\mathbf{z})$. When the Cox PH model is used then $g(\boldsymbol{\beta}'\mathbf{z}) = e^{\boldsymbol{\beta}'\mathbf{z}}$, and the hazard and survival functions have the following form:

$$h(x|\mathbf{z};\boldsymbol{\beta}) = h_0(x)e^{\boldsymbol{\beta}'\mathbf{z}}, \tag{4.19}$$

and

$$S(x|\mathbf{z};\boldsymbol{\beta}) = S_0(x)e^{\boldsymbol{\beta}'\mathbf{z}}. \tag{4.20}$$

As we can see from the above equation, the Cox model reduces to the baseline hazard function when all the covariates are 0. Cox proportional hazard model is termed semiparametric as $h_0(x)$ is an arbitrary baseline hazard, but $g(\boldsymbol{\beta}'\mathbf{z})$ is specified. Therefore, if the proportional hazards assumption holds, it is possible to estimate the effect of parameters without any consideration of the hazard function. Although, no assumption is made about the probability

distribution of the hazard, it is assumed that the hazard ratio does not depend on time i.e., the risk of experiencing an event at a particular point in time in one group is a multiplication of the risk in the other group, and it will stay constant at any other time [31]. An important feature of the Cox model is that the covariates are not dependent on time (e.g. sex, place of birth, etc.). Let's see what happened to the PH model assumption if a length-biased function is considered. By, already known from Chapter 2, relationship between the density, survival, and hazard functions, and by 4.19 and 4.20, we have the following expression for the unbiased hazard function:

$$h_U(x|\mathbf{z}) = \frac{f_U(x|\mathbf{z})}{S_U(x|\mathbf{z})} \tag{4.21}$$

In the presence of length bias (and by the definition of a length-biased density), the equation 4.21 can be re-written as:

$$\begin{aligned} h_{LB}(x|\mathbf{z}) = \frac{f_{LB}(x|\mathbf{z})}{S_{LB}(x|\mathbf{z})} &= \frac{\frac{xf_U(x|\mathbf{z})}{\mu(\mathbf{z})}}{\int_x^\infty \frac{wf_U(w|\mathbf{z})}{\mu(\mathbf{z})}dw} \\ &= \frac{xh_0(x)e^{\boldsymbol{\beta}'\mathbf{z}}S_0(x)^{e^{\boldsymbol{\beta}'\mathbf{z}}}}{\int_x^\infty wh_0(w)e^{\boldsymbol{\beta}'\mathbf{z}}S_0(w)^{\boldsymbol{\beta}'\mathbf{z}}dw}. \end{aligned} \tag{4.22}$$

In order for the assumption of proportionality to be satisfied,

$$h_{LB}(x|\mathbf{z}) = h_0(x)c(\beta'\mathbf{z}), \tag{4.23}$$

where $c(\beta'\mathbf{z})$ is some function depending on covariates [9]. In other words,

$$xe^{\boldsymbol{\beta}'\mathbf{z}}S_0(x)^{e^{\boldsymbol{\beta}'\mathbf{z}}} = c(\beta, \mathbf{z})\int_x^\infty wh_0(w)e^{\boldsymbol{\beta}'\mathbf{z}}S_0(x)^{\boldsymbol{\beta}'\mathbf{z}}dw, \tag{4.24}$$

which clearly is not a case (taking the derivative on both side would make it even clearer). Therefore, even if the assumption of the PH models holds in the unbiased case, it cannot be extended to the length-biased sampling scenario.

The basic difference between AFT and PH models is the quantities compared: AFT models compare survival functions while PH models compare hazard functions. Also, the effect of covariates in AFT models is proportional with respect to time while in PH models multiplicative with respect to the hazard function. The popularity of the Cox model comes from its robustness i.e., from the fact that if the correct parametric model is, for example, Weibull, then the

estimates of regression coefficients, hazard ratios, and adjusted survival curve from the Cox model will be very close to those obtained from the parametric model without specifying the baseline function. If specifying the correct distribution is problematic, the Cox model would be a safe choice provided the assumption of PH are met.

It is worth noting that the exponential and Weibull distributions satisfy the assumptions of both the AFT and PH models i.e., if the AFT assumption holds then the PH assumption also holds (and vice versa). In the case of the Weibull distribution, the only requirement for this property to hold is for the shape parameter $\alpha$ to be constant over different values of covariates [29]. Recall that the hazard function for the Weibull distribution with parameters $(\alpha, \lambda)$ is $h(x) = \alpha\lambda x^{\alpha-1}$, and the Weibull proportional hazard model can be written as:

$$h(x|\mathbf{z}; \boldsymbol{\beta}) = \alpha\lambda x^{\alpha-1} e^{\boldsymbol{\beta}'\mathbf{z}} = \alpha x^{\alpha-1}(\lambda e^{\boldsymbol{\beta}'\mathbf{z}}), \tag{4.25}$$

which is again a hazard function of the Weibull distribution with a scale parameter $\lambda e^{\boldsymbol{\beta}'\mathbf{z}}$ and shape parameter $\alpha$. Therefore, the Weibull family with fixed $\alpha$ possesses PH property. This shows that the covariates in the model change the scale parameter of the distribution, while the shape parameter remains constant. So, in Weibull PH model, $h(x) = \lambda\alpha x^{\alpha-1}$, where $\lambda = e^{\beta'\mathbf{z}}$. Let $Z$ be a simple covariate with $Z = 1$ if a subject is male and $Z = 0$ if female. The hazard ratio (HR) of $Z = 1$ vs $Z = 0$ can be then expressed as follows:

$$HR = \frac{e^{\beta_0+\beta_1}\alpha x^{\alpha-1}}{e^{\beta_0}\alpha x^{\alpha-1}} = e^{\beta_1}. \tag{4.26}$$

If $HR > 1$, we would say (similarly to odds ratios in the logistic models) that the risk of having an outcome variable (for example smoking) for men is $e^{\beta_1}$ times higher than that of women.

Let's now look at the Weibull distribution when AFT model is used. The survival function for the Weibull $(\alpha, \lambda)$ distribution is $S(x) = e^{-\lambda x^\alpha}$. When solving for $x$ ($x$ is a time variable here), we obtain the following:

$$x = (-\log(S(x)))^{1/\alpha} \frac{1}{\lambda^{\frac{1}{\alpha}}}. \tag{4.27}$$

By parameterizing the latter factor as $\frac{1}{\lambda^{\frac{1}{\alpha}}} = e^{\boldsymbol{\tau}'\mathbf{z}}$, we can express $x$ as

$$x = (-\log(S(x)))^{1/\alpha} \mathbf{e}^{\boldsymbol{\tau}'\mathbf{z}}. \tag{4.28}$$

Therefore, for each fixed value of $S(x)$, $e^{\boldsymbol{\tau}'\mathbf{z}}$ is a time scaling factor, and Weibull distribution satisfies the assumption of the AFT models [36]. Based on 4.28, the acceleration factor $\gamma$ (a quantity that characterizes the AFT models) for the above men/women example can be calculated as follows:

$$\gamma = \frac{(-\log(S(x)))^{1/\alpha} e^{\tau_0+\tau_1}}{(-\log(S(x)))^{1/\alpha} e^{\tau_0}} = e^{\tau_1}. \tag{4.29}$$

The interpretation of $\gamma$ would be that the survival time at any time point $x$ is increased by the factor $e^{\tau_1}$ for men compared to women. For a Weibull distribution (and also exponential), once we obtain a regression coefficient for the PH model, we can easily calculate the regression coefficient for the AFT model (and vice versa) by means of the following relationship between the coefficients in the two models: $\beta_i = -\tau_i\alpha$ where $\beta_i$ and $\tau_i$ are regression coefficients in the PH and AFT models respectively, and $\alpha$ is the shape parameter of Weibull distribution.

In terms of survival and hazard functions for a Weibull distribution, we have the following relationship based on the relationship between its PH and AFT representation: if $e^{\beta'\mathbf{z}} > 1 \Longrightarrow \beta > 0$, the hazard function in PH models increases (and survival therefore decreases), while in AFT models, the time to event increases, and thus survival increases and the hazard decreases.

Because of its flexibility, Weibull distribution is one of the most widely used distribution in survival analysis.

## 4.2 Maximum likelihood estimates in the presence of covariates

The estimation procedures described in details in Chapter 2 can be further extended to obtain the estimates of parameters in the presence of covariates. In this section, we will derive maximum likelihood functions for unbiased, length-biased, and a combination of both types of data cases in the presence of covariates.

### 4.2.1 MLE in unbiased sampling with covariates

The maximum likelihood estimation procedure is quite straightforward in the case when data come from unbiased sampling such as incident studies. In this case, the survival times are a random sample from the population subjected to possible censoring and conditioned upon the value of the vector of covariates $\mathbf{Z}$. Since the data are unbiased, the distribution of covariates does not depend on the parameters of interest, i.e. the covariates can be conditioned out without a concern about a loss of information about the survival distribution [8]. Also, censoring is assumed to be independent of the event times and therefore regarded as being noninformative.

Let's denote the likelihood function for incident data by $L_{Inc}$ and let $X_1, \ldots, X_n$ be a random sample of survival times with a common cumulative distribution function $F_U(x)$ and a probability distribution function $f_U(x)$. Also, let $\boldsymbol{\theta}$ be a vector of parameters associated with $F_U(x)$ and $\mathbf{z_i}$ be a vector of covariates for the $i$th subject. Then, the corresponding likelihood function has a form parallel to the likelihood function when no covariates are considered:

$$L_{Inc}(\boldsymbol{\theta}) = \prod f_U(x|\mathbf{z_i};\boldsymbol{\theta})^{\delta_i} S_U(x|\mathbf{z_i};\boldsymbol{\theta})^{(1-\delta_i)}. \tag{4.30}$$

Here, $\delta_i$ is the indicator of whether an event ($\delta_i = 1$) or censoring ($\delta_i = 0$) occurred. The estimation of the parameters $\boldsymbol{\theta}$ of $f_U(x)$ follows the usual maximization procedure based on partial derivatives of the log-likelihood of $L_{Inc}$. The obtained MLE $\hat{\boldsymbol{\theta}}$ are the unbiased estimates of the true population parameters $\boldsymbol{\theta}$ and $\hat{S}_{Inc}$ is an unbiased MLE of the survival function in the incident population.

### 4.2.2 MLE in length-biased sampling with covariates

The situation becomes slightly more complicated in the case of prevalent data, in which, due to length-biased sampling, survival times can no longer be considered as a random sample from the target population. The lack of randomness, in this case, may introduce bias into the distribution of covariates since they are now related to the long-term survivors. For example, if longer times are most likely to be sampled and women tend to live longer, the effect of gender on survival times will be weighted by the effect of being a women. Therefore, in the

case of prevalent data, conditioning on covariates is not recommended and a more appropriate, unconditional approach, should be used instead [8]. In contrast to the conditional approach, in the unconditional approach, we assume that the covariates effect is related to the lifetime distribution, and therefore depends on the parameters of the distribution of failure times. An unconditional approach incorporates additional information carried out in the distribution of covariates to the likelihood function and results in more efficient estimates [7]. It also helps to recover the underlying unbiased distribution of survival times in a prevalent cohort [9].

Following [8], let $X_i$ and $V_i$ represent the total failure time and the total censoring time for subject $i$ in a length-biased sample with right censoring. Recall from the previous chapters that under this scenario, the observations are independent, but $X$ and $V$ are not. They share a common truncation time, and therefore the censoring is informative. If we let $T_i$ to be a truncation time for observation $i$, then $X_i = T_i + R_i$ and $V_i = T_i + C_i$, where $R_i$ and $C_i$ are the residual lifetime and residual censoring time, respectively. In our scenario, we also assume independence between $C_i$ and $(T_i, R_i)$. Let $f$ be a joint density function of the observed vector $\vec{w} = (t_i, r_i \wedge c_i, \mathbf{z_i}, \delta_i)$, where $\delta_i = 1$ when the failure time is observed. We would like to estimate the unbiased distribution function of survival as well as the covariates effect on the survival times. Let $F_{LB}(x, \mathbf{z})$ denote the joint distribution of the lifetimes and covariates coming from the unbiased joint distribution $F_U(x, \mathbf{z})$ with parameter vector $\boldsymbol{\theta}$. In addition, let $F_Z(\mathbf{z})$ and $F_B(\mathbf{z})$ denote the distribution of covariates under unbiased and biased sampling, respectively.

In Chapter 2, we showed that the maximum likelihood function for length-biased data without covariates has the following form:

$$L_{LB}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left( \frac{f_U(x; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} \right)^{\delta_i} \left( \int_{w \geq v_i}^{\infty} \frac{f_U(w; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} dw \right)^{1-\delta_i}, \tag{4.31}$$

where $\delta_i$ is an indicator of death or censoring for $i$th observation and $\mu(\boldsymbol{\theta})$ is the mean survival time in the unbiased population.

The conditional approach simply extends the above likelihood in the following way:

$$L_{Cond}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left( \frac{f_U(x|\mathbf{z_i}; \boldsymbol{\theta})}{\mu(\mathbf{z_i}; \boldsymbol{\theta})} \right)^{\delta_i} \left( \int_{w \geq v_i}^{\infty} \frac{f_U(w|\mathbf{z_i}; \boldsymbol{\theta})}{\mu(\mathbf{z_i}; \boldsymbol{\theta})} dw \right)^{1-\delta_i}, \tag{4.32}$$

with $\mu(\mathbf{z_i};\boldsymbol{\theta}) = E(X|\mathbf{z_i})$ being the mean value of survival time for a given value of covariate $\mathbf{z}$.

In order to incorporate the information in covariates to a likelihood function, we have to consider the following **joint** likelihood:

$$L_{Joint}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_U(x_i, \mathbf{z_i}|\boldsymbol{\theta})^{\delta_i} (\int_{w \geq v_i}^{\infty} f_U(w_i, \mathbf{z_i}|\boldsymbol{\theta}) dw)^{1-\delta_i}. \tag{4.33}$$

Next, using the relation between the joint and conditional density functions, we can write the above likelihood for each $i$ as follows:

$$\begin{aligned} L_{Joint,i}(\boldsymbol{\theta}) &= f_U(x_i, \mathbf{z_i}|\boldsymbol{\theta})^{\delta_i} \left( \int_{w \geq v_i}^{\infty} f_U(w_i, \mathbf{z_i}|\boldsymbol{\theta}) dw \right)^{1-\delta_i} \\ &= f_u(x_i|\mathbf{z_i}, \boldsymbol{\theta})^{\delta_i} \left( \int_{w \geq v_i}^{\infty} f_U(w_i|\mathbf{z_i}, \boldsymbol{\theta}) dw \right)^{1-\delta_i} f_B(\mathbf{z_i};\boldsymbol{\theta}). \end{aligned} \tag{4.34}$$

It can be easily seen that the above likelihood is equivalent to

$$L_{Joint}(\boldsymbol{\theta}) = L_{Cond}(\boldsymbol{\theta}) \prod_{i=1}^{n} f_B(\mathbf{z_i};\boldsymbol{\theta}), \tag{4.35}$$

where $f_B$ is the biased distribution function of covariates.

Let's now look at how the biased distribution of covariates is derived. Denote by $X$ the survival time and by $T$ the truncation time. All of the assumptions discussed above hold here as well. We also know that $X$ is observed only if $X > T$. In this setting, the biased distribution of covariates is derived as follows:

$$f_B(\mathbf{z};\boldsymbol{\theta}) = f(\mathbf{z}|X > T;\boldsymbol{\theta}) = \frac{P(X \geq T|\mathbf{z};\boldsymbol{\theta}) f_Z(\mathbf{z})}{P(X \geq T;\boldsymbol{\theta})}, \tag{4.36}$$

where $f_Z(\mathbf{z})$ is the unbiased distribution of covariates. Since $X$ and $T$ are independent, we have

$$P(X \geq T|\mathbf{z};\boldsymbol{\theta}) = \int_0^{\infty} \int_0^{x} f_U(x|\mathbf{z};\boldsymbol{\theta}) f_T(t) dt dx. \tag{4.37}$$

Furthermore, because of the stationarity assumption, the truncation time $T$ has uniform, independent of covariates, distribution $f_T(t)$. Under this assumption, the above equation reduces to

$$P(X \geq T|z;\boldsymbol{\theta}) \propto \int_0^{\infty} \int_0^{u} f_U(x|z;\boldsymbol{\theta}) dt dx = \int_0^{\infty} x f_U(x|z;\boldsymbol{\theta}) dx = \mu(z;\boldsymbol{\theta}), \tag{4.38}$$

and the length-biased density of covariates becomes

$$f_B(z;\boldsymbol{\theta}) = f(z|X > T;\boldsymbol{\theta}) = \frac{P(X \geq T|z;\boldsymbol{\theta})f_Z(z)}{\int_z P(X \geq T|z)f_Z(z)dz} \\ = \frac{\mu(z;\boldsymbol{\theta})f_Z(z)}{\int_z \mu(z;\boldsymbol{\theta})f_Z(z)dz} = \frac{\mu(z;\boldsymbol{\theta})f_Z(z)}{\mu(\boldsymbol{\theta})}, \tag{4.39}$$

where $\mu(\boldsymbol{\theta}) = E(E(X|Z;\boldsymbol{\theta})) = E(X)$ is the marginal mean lifetime of the incident population [7].

Finally, since

$$\mu(\mathbf{z};\boldsymbol{\theta}) = \int_0^\infty S_U(x|\mathbf{z};\boldsymbol{\beta})dx = \int_0^\infty S_0(xe^{\boldsymbol{\beta}'\mathbf{z}})dx = e^{\boldsymbol{\beta}'\mathbf{z}} \int_0^\infty S_0(w)dw \\ = e^{\boldsymbol{\beta}'\mathbf{z}}\mu(0) = e^{\boldsymbol{\beta}'\mathbf{z}}\mu(0), \tag{4.40}$$

the density can be written as follows:

$$f_B(\mathbf{z};\boldsymbol{\theta}) = \frac{e^{\boldsymbol{\beta}'\mathbf{z}}\mu(0)f_Z(\mathbf{z})}{\int_{\mathbf{z}} e^{\boldsymbol{\beta}'\mathbf{z}}\mu(0)f_Z(\mathbf{z})dz} \\ = \frac{e^{\boldsymbol{\beta}'\mathbf{z}}f_Z(\mathbf{z})}{E(e^{\boldsymbol{\beta}'\mathbf{z}})} \tag{4.41}$$

As further shown in [7], the joint likelihood function for the prevalent population can be ultimately written as

$$L_{Joint}(\boldsymbol{\theta}) \propto L_C(\boldsymbol{\theta})L_Z(\boldsymbol{\theta}), \tag{4.42}$$

where $L_Z(\boldsymbol{\theta})$ is the likelihood function of covariates only.

After some algebraic manipulation and elimination of terms not contributing to the likelihood function, we have

$$L_{Joint}(\boldsymbol{\theta}) \propto \left(\frac{f_U(x_i|\mathbf{z_i};\boldsymbol{\theta})}{\mu(\mathbf{z},\boldsymbol{\theta})}\right)^{\delta_i} \left(\int_{w \geq v_i} \frac{f_U(w|\mathbf{z_i};\boldsymbol{\theta})}{\mathbf{z},\mu(\boldsymbol{\theta})}dw\right)^{1-\delta_i} \frac{e^{\boldsymbol{\beta}'\mathbf{z}}f_Z(\mathbf{z})}{E(e^{\boldsymbol{\beta}'\mathbf{z}})}. \tag{4.43}$$

One can notice that the "only" difference between the conditional and joint approach is the expression for $\mu$: the mean in the joint approach is an overall unconditional mean lifetime of the unbiased population while in the conditional approach it depends on value of covariate $\mathbf{z}$.

The MLEs obtained from the joint likelihood was proven to have better efficiency than the one from the conditional likelihood in the presence of length bias. Its asymptotic properties were extensively studied by Ashgarian [5].

### 4.2.3 MLE for combined incident and prevalent data with covariates

In this section, we will extend the combined maximum likelihood function derived earlier in Chapter 2 to include covariates. Recall that the observations in the combined dataset are assumed to be independent and come from the same underlying distribution.

Let $L_{Inc}$ be the likelihood function for the incident sample given by 4.30, $L_{Prev}$ be the likelihood function of prevalent data given by 4.43, and let $f_Z(\mathbf{z})$ define the unbiased distribution function of covariates. Then the likelihood function for combined prevalent and incident data $L_{Comb}(\boldsymbol{\theta})$ can be written as follows:

$$L_{Comb}(\boldsymbol{\theta}) = L_{Inc}(\boldsymbol{\theta})^{\xi_i} \times L_{Prev}(\boldsymbol{\theta})^{1-\xi_i}, \tag{4.44}$$

where $\xi_i = 1$ if an observation comes from the incident sample, and 0 otherwise. Equivalently, we can write the combined likelihood as

$$\begin{aligned} L_{Comb}(\boldsymbol{\theta}) = \prod_{i=1}^{n} &\left[f_U(x_i|\mathbf{z_i})^{\delta_i} S_U(x_i|\mathbf{z_i})^{1-\delta_i} f_Z(\mathbf{z_i})\right]^{\xi_i} \\ &\times \left[\frac{1}{\mu(\boldsymbol{\theta})} f_U(x_i|\mathbf{z_i})^{\delta_i} S_U(x_i|\mathbf{z_i})^{1-\delta_i} f_Z(\mathbf{z_i})\right]^{1-\xi_i}, \end{aligned} \tag{4.45}$$

where $\xi_i$ is defined as above and $\delta_i$ is the indicator of whether an event or censoring occurred. Simplifying 4.45, we obtain:

$$L_{Comb}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left(\frac{1}{\mu(\boldsymbol{\theta})}\right)^{1-\xi_i} f_U(x_i)^{\delta_i} S_U(x_i|\mathbf{z_i})^{1-\delta_i} f_Z(\mathbf{z_i}). \tag{4.46}$$

To estimate the survival function by means of the combined likelihood, we utilize all of the information contained in both prevalent and incident data, as well as the information about the regression parameters contained in the sampling distribution of the covariates; therefore, the MLEs obtained from the above likelihood are expected to be more efficient than the one from the incident and prevalent sample alone.

### 4.2.4 MLE for covariates only

It is interesting to note that in the extreme case when some of the data are lost, but information on covariates and an indication of whether a given observation is a prevalent or incident case

is available, we could still estimate the regression coefficients from the likelihood function of the covariates only. It was shown with a simulation study that although the variance of the estimates increases in this case, the bias remains close to the case when the full data are available [8].

The likelihood function for covariates from the combined (prevalent and incident) data would be:

$$L_{Cov}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_U(\mathbf{z_i})^{\xi_i} f_B(\mathbf{z_i}, \boldsymbol{\theta})^{1-\xi_i}, \tag{4.47}$$

where $\xi_i$ is indicator of whether an observation is an incident or prevalent case, and $\boldsymbol{\theta}$ is a vector of parameters to be estimated. $f_U$ and $f_B$ represent unbiased and biased distribution functions, respectively. For AFT models, we can re-write the equation 4.47 with the biased distribution of covariates defined explicitly using equation 4.41:

$$L_{Cov}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_U(\mathbf{z_i})^{\xi_i} \left[\frac{e^{\boldsymbol{\beta}'\mathbf{z}} f_Z(\mathbf{z})}{E(e^{\boldsymbol{\beta}'\mathbf{z}})}\right]^{1-\xi_i}, \tag{4.48}$$

where $\mu(\boldsymbol{\theta})$ is the overall mean lifetime in the incidence population. Besides the covariate distribution in the incident population, based on equation 4.48, one can notice that the sampling distribution of covariates in the length-biased scenario in AFT models $(f_B(\mathbf{z_i}, \boldsymbol{\theta}))$ does not depends on the lifetime distribution, but only on the regression parameters $\boldsymbol{\theta}$. Therefore, the effect of covariates on survival can be estimated having only some information on samples' characteristics (age, sex, etc.). For example, supposed that a researcher is interested in the length of time it takes a recent graduate from the Department of Mathematics and Statistics to find a job. He, therefore, obtains a list of students who graduated one semester earlier and still do not have a job, and decides to follow them up until an event (finding a job) occurs. However, for some reason, he must terminate the research, and the only information he is able to obtain is a program from which a student graduated: pure mathematics, or applied mathematics/ statistics. Assuming that the number of student graduating in both programs was approximately the same, the researcher finds out that more math than stats' students are still looking for a job, which would imply that the length of time an average stats' graduate looks for a job is shorter than the one of a math student. Since the proportion of stats' students in

the sample is smaller, it means that they must have found a job within the six months from graduation (and before the study began). Therefore, just based on the covariate information (the student's program), we could obtain some information about the survival time (the length of time until the first job).

The situation is not that simple in the PH models. For simplicity, let's use again an example with a Weibull distribution. Recall that the hazard function for the unbiased Weibull PH model in the presence of covariates is given by

$$h(x|\mathbf{z};\boldsymbol{\theta}) = e^{\boldsymbol{\beta}'\mathbf{z}}\alpha\lambda^{\alpha}x^{\alpha-1} = \alpha(e^{\frac{\boldsymbol{\beta}'\mathbf{z}}{\alpha}})^{\alpha}x^{\alpha-1}, \tag{4.49}$$

which is again the hazard function of Weibull $(\alpha, e^{\frac{\boldsymbol{\beta}'\mathbf{z}}{\alpha}})$ with mean

$$\mu(\mathbf{z};\boldsymbol{\theta}) = \frac{e^{-\frac{\boldsymbol{\beta}'\mathbf{z}}{\alpha}}\Gamma(1+1/\alpha)}{\lambda} \tag{4.50}$$

(The mean of a Weibull $(\alpha, \lambda)$ is $\mu(\alpha, \lambda) = \Gamma(1+1/\alpha)/\lambda$ )

It can be shown, that the sampling distribution of the covariates in this setting is

$$f_B(\mathbf{z};\alpha,\lambda,\boldsymbol{\beta}) = \frac{e^{-\frac{\boldsymbol{\beta}'\mathbf{z}}{\alpha}}f_{\mathbf{z}}(\mathbf{z})}{\int_{\mathbf{z}} e^{-\frac{\boldsymbol{\beta}'\mathbf{z}}{\alpha}}f_{\mathbf{z}}(\mathbf{z})d\mathbf{z}}. \tag{4.51}$$

Therefore, in PH models, the mean lifetime in the (unbiased) population cannot be as nicely factored out like it is done in the case of AFT models, and, in consequence, the regression parameters cannot be simply estimated by just a distribution of covariates since the shape parameter of the hazard "contaminates" the covariate distribution with length-biased data in PH model [9].

Once we decide what model (and possibly distribution) describe best our data, we can proceed to constructing a likelihood function to estimate the parameters of interest. When unbiased and length-biased data are treated separately, the likelihood function is quite straight forward to obtain and does not require any extra assumptions (as discussed above); however, with the likelihood function for combined data, we must make sure that both the prevalent and incident cases come from the same distribution. The method to test this assumption will be presented in the next chapter.

# Chapter 5

# Assessing the equality of incident and prevalent distributions

The main objective of this thesis is to develop a maximum likelihood estimator of the survival function from a combined incident and prevalent data. In order to meet this objective, we first have to assess if both types of data come from the same population, i.e. if the underlying distribution functions (and therefore the survival functions) are the same in both populations. If this assumption does not hold, the estimates based on the combined data would be invalid and our approach not applicable. To test whether two independent samples come from the same population, we will develop a Smirnov type statistics based on the two-sample Kolmogorov-Smirnov (KS) goodness-of-fit (Gof) test developed in 1939 by Smirnov.

The goal of this section is, first, to describe the KS test and compare it briefly to two other popular GoF tests, Anderson-Darling (AD) and Chi-Squared to justify our choice, and next to introduce an adaptation of the Smirnov test that meets our objectives.

## 5.1 Kolmogorov-Smirnov goodness-of-fit test

Goodness of fit tests are essentially based on either of two distributions: the cumulative distribution function ($cdf$) or the probability density function ($pdf$). While the Chi-squared test is

based on the $pdf$, both the AD and the KS GoF tests use the cdf approach and hence belong to the class of "distance tests". The KS test has the advantage of making no assumption about the distribution of data: it is a nonparametric, distribution free alternative to the two-sample t-test applied when the assumptions of parametric tests are violated. The main advantage of the two-sided KS test over similar GoF tests is its sensitivity to differences in both location and shape of the empirical cumulative distribution functions, as well as its consistency against all types of differences that may exist between the two distribution functions: a difference with respect to location/central tendency, variability, skewness and kurtosis [**?**]. In contrast to the KS test, the Anderson-Darling test, a modification of the KS test, puts more weight on the tail of a distribution and for this reason is recommended when a more sensitive comparison for the bigger values (or later times) is required. Unlike the Chi-squared test, both the AD and KS tests perform well in the case of small sample sizes and their performance does not depend on how the data are categorized. In addition, the KS test is easier to apply than the AD test since its critical region does not depend on the distribution used. The main disadvantage of the KS test over the Chi-squared test is its inability to make $k$-sample comparisons and accommodate nominal data. Nonetheless, none of these characteristics are relevant to our setting, and therefore the KS test seems to be the most feasible nonparametric GoF test choice.

The KS tests refer to a family of GoF tests that may be used in one- or two- sample situation. To be more precise, statistics which are functions of the distance between the empirical distribution function and the hypothesized distribution function are known as Kolmogorov-type statistics, whereas statistics which are functions of the vertical distance between two empirical distribution functions are known as Smirnov-type statistics [35].

Since we will be using a KS type of test, we will discuss now the basic properties of the KS tests. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be i.i.d random variables with $cdf$s $F(x)$ and $G(x)$, respectively. Also, let $F_m(x) = P_m(X \leq x) = \frac{1}{m}\sum_{i=1}^{m} I(X_i \leq x)$ and $G_n(x) = P_n(Y \leq x) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq x)$ be the empirical $cdf$s defined on the same interval. By the law of large numbers, $F_m(x)$ and $G_n(x)$ converge in distribution to $F(x)$ and $G(x)$. Consequently,

$$\sup_{x \in \mathbb{R}} |F_m(x) - F(x)| \to 0 \tag{5.1}$$

and

$$\sup_{x\in\mathbb{R}} |G_n(x) - G(x)| \to 0. \tag{5.2}$$

The central observation of the KS test is that distribution of the supremum does not depend on the distribution of the sample. The proof of this property was developed by Panchenko and is presented in details below [39]:

**Theorem 5.1.1.** *If $F(x)$ is a continuous cumulative distribution function then the distribution of*

$$\sup_{x\in\mathbb{R}} |F_m(x) - F(x)| \tag{5.3}$$

*does not depend on $F$.*

**Proof.** Define the inverse of $F(x)$ by

$$F^{-1}(y) = \min(x : F(x) \geq y). \tag{5.4}$$

Using the following change of variables: $y = F(x)$ and $x = F^{-1}(y)$, we can show that

$$P(\sup_{x\in\mathbb{R}} |F_m(x) - F(x)| \leq t) = P(\sup_{0\leq y\leq 1} |F_m(F_{-1}(y)) - y| \leq t). \tag{5.5}$$

By definition of the empirical $c.d.f$, we obtain

$$F_n(F^{-1}(y)) = \frac{1}{m}\sum_{i=1}^{m} I(X_i \leq F^{-1}(y)) = \frac{1}{m}\sum_{i=1}^{m} I(F(X_i) \leq y). \tag{5.6}$$

Therefore,

$$P(\sup_{0\leq y\leq 1} |F_m(F_{-1}(y)) - y| \leq t) = P\left(\sup_{0\leq y\leq 1} \left|\frac{1}{m}\sum_{i=1}^{m} I(F(X_i) \leq y) - y\right| \leq t\right). \tag{5.7}$$

Now, because the following holds for the c.d.f. of $F(X_1)$:

$$P(F(X_1) \leq t) = P(X_1 \leq F_{-1}(t)) = F(F_{-1}(t)) = t, \tag{5.8}$$

the distribution of $F(X_i)$ is uniform on the interval $[0, 1]$. Finally, the random variables $U_i = F(X_i)$ (for $i \leq n$) are independent uniform $r.v$s. It follows from above that

$$P(\sup_{x\in\mathbb{R}} |F_m(x) - F(x)| \leq t) = P\left(\sup_{0\leq y\leq 1} \left|\frac{1}{m}\sum_{i=1}^{m} I(U_i \leq y) - y\right| \leq t\right) \tag{5.9}$$

which is independent of $F$ [39].

In the two-sample Smirnov test, the null and the alternative hypotheses can be stated as follows:

$$H_0 : F_m(x) = G_n(x) \text{ for all } x, \tag{5.10}$$

and

$$H_A : F_m(x) \neq F_n(x). \tag{5.11}$$

Theorem 5.1.2 defines test statistic for the KS test.

**Theorem 5.1.2.** *Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be i.i.d. r.vs with a common continuous c.d.f. and let $F_m(x)$ and $G_n(x)$ be empirical c.d.fs of $X$'s and $Y$'s, respectively. Furthermore, let*

$$D_{m,n} = \left(\sqrt{\frac{mn}{m+n}}\right) \sup_{x \in \mathbb{R}} |F_m(x) - G_n(x)|. \tag{5.12}$$

*Then,*

$$\lim_{m,n \to \infty} P\left(D_{m,n} \leq t\right) = Q(t), \tag{5.13}$$

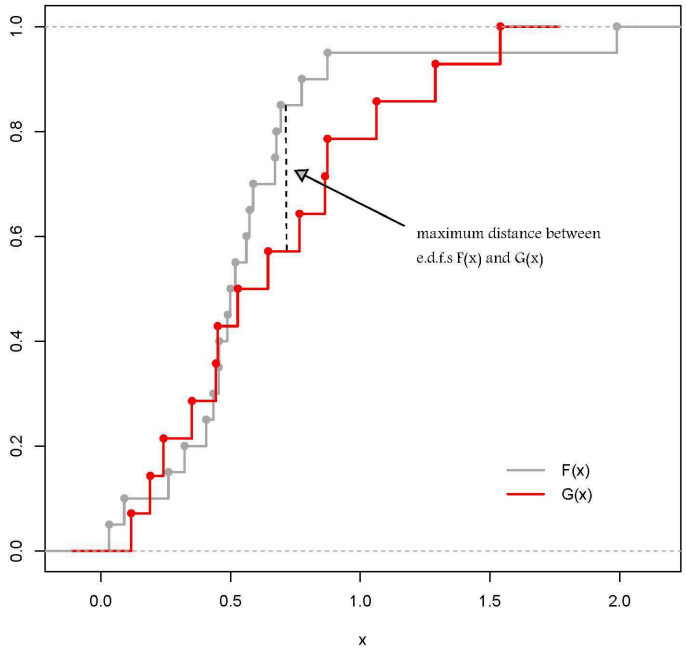
*where $Q(t)$ is the c.d.f. of the Kolmogorov-Smirnov distribution given by*

$$Q(x) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2x^2}. \tag{5.14}$$

The sampling distribution of $D_{m,n}$ is known and depends only on the size of the samples. The probabilities associated with the occurrence of values as large as an observed under the null hypothesis have been tabulated for each $n$ and we can find the threshold values directly from the tables.

Note that in survival analysis $S(x)$ is used instead of $F(x)$. For the purpose of testing, however, $F(x)$ and $S(x)$ are equivalent ($F(x) = 1 - S(x)$), and the change does not affect estimation process. Therefore, $F_m(x)$ and $G_n(x)$ can be easily replaced by $1 - S_m(x)$ and $1 - S_n(x)$ in the above above equations, and $D_{m,n}$ can be calculated using survival functions instead of $cdf$s as it will be done in our simulations. Figure 5.1 illustrate how the distance between two $edf$s is measured. Also, in our case, the p-values for the test cannot be used directly from the tables (as explained in the next paragraphs), and simulations have to be used to roughly estimates the probabilities of rejecting the null hypothesis.

Figure 5.1: Maximum distance ($D_{m,n}$- statistic) between two $edf$s: $F(x)$ and $G(x)$




Let's now look at how to adapt the Smirnov test to our scenario. The Smirnov type statistics rely on empirical distribution function of the underlying data, the nonparametric maximum likelihood estimator (NPMLE) that assumes no censoring or truncation, and therefore is based on a random sample of complete observations. For this reason, we cannot use the Smirnov test directly as, in our case, full information on survival time is not available, and consequently the empirical survival function is not an appropriate estimator. Instead, we have "equivalents" of the empirical estimators: the KM estimator, the NPMLE for incident cases with right censoring, and the Vardi's NPMLE that adjusts for length bias in prevalent data. To suit the properties of our samples, we will develop an "ad hoc" Smirnov test that will quantify the discrepancy between the distributions using a distance statistic.

Finding the distribution of the statistic ($D_{m,n}$) under the null hypothesis, estimating the $p$-value and the power of our test, while taking into account sample size and censoring, will be done by simulating data from a known population distribution. In our case, we will set up

our simulations to reflect, as close as possible, the incident and prevalent samples from the CSHA data. The simulations will help us to determine behavior of the proposed statistic in the particular scenario when the data is subjected to censoring.

Another reason to perform simulations is "imperfectness" of our NPMLEs. In theory, since the KM estimator is applied to incident sample and the Vardi's NPMLE corrects for length bias in the prevalent data, the two distributions should be very close and describe the same population. However, in practice, both estimators have limitations that introduce uncertainty to estimation of parameters for our type of survival data. For example, since the KM is undefined after the largest observed event-time and its efficiency decreases with increased amount of censoring in a sample, the survival estimates for these times are not as reliable as they are for the earlier times when the risk set (i.e. a number of subjects exposed) is larger. On the other hand, the Vardi curve, is more variable at the left-hand tail with a tendency to underestimate the survival early on for finite samples.

In the next chapter, we will present the algorithms used to generate the incident and prevalent samples, obtain the maximum distances between the samples, as well as to investigate behavior of the proposed statistic in the context of our setting (CSHA survival data). In the Applications chapter, we will apply the test to compare the nonparametric MLEs of the survival functions obtained from the incident and prevalent populations using the Kaplan-Meier estimator and the NPMLE proposed by Vardi, respectively. Based on the obtained results, we will decide whether or not estimation of the survival function using combined incident and prevalent data is feasible.

# Chapter 6

# Simulations

In statistics, simulations (computer intensive procedures) are often used to estimate accuracy measures and provide an empirical estimation of the sampling distribution of parameters that could not be obtained from a single study. These techniques also enable to test hypotheses and validate statistical methods. They are in many ways easy to use and generalize to situations where traditional methods based on the theory of normal distribution cannot be applied and close form of an estimate is not available. Initiated by Efron in 1979, a basic bootstrap approach uses Monte Carlo sampling to generate an empirical estimate $\hat{\theta}$ from the empirical distribution $\hat{F}$ where $\hat{\theta}$ is an estimate of the true parameters $\theta$ (mean, median, correlation, and so forth) of an unknown probability distribution $F$, and $\hat{F}$ is an empirical estimate of the true distribution. Monte Carlo sampling builds an estimate of the sampling distribution by randomly drawing with replacement a large number of samples of size $n$ from a population, and calculating for each one the associated value of the statistic $\hat{\theta}$. The relative frequency distribution of these $\hat{\theta}$ values is an estimate of the sampling distribution of the statistic of interest. Larger the number of samples of size $n$, the more accurate the relative frequency distribution of these estimates will be [20].

Based on a desired accuracy of the estimates, the number of simulations, $B$, can be calculated as follows:

$$B = \left( \frac{Z_{1-(\alpha/2)}\sigma}{\delta} \right)^2, \tag{6.1}$$

where $\delta$ is the specified level of accuracy of the estimate, $Z_{1-(\alpha/2)}$ is the $1-(\alpha/2)$ quantile of the standard normal distribution, and $\sigma$ is the standard deviation of the parameter of interest [12].

Bootstrapping from uncensored data is relatively simple. The situation becomes more complicated when censoring and/or truncation has to be taken into account. Efron (1981) showed how to resample a right-censored data when censoring is assumed to be random or fixed. In this scenario, we obtain two independent random samples: for a lifetime variable $T_i$ and a censoring variable $W_i$. We define $X_i$ as a $\min(T_i, W_i)$. The obtained data will constitute of $n$ pairs $(X_i, D_i)$ with $D_i = 1$ if $X_i = T_i$ and $D_i = 0$ when $X_i = W_i$. In the case when censoring times $w_i, w_2, \ldots, w_n$ are assumed to be fixed, we define $X_i$ as a $\min(T_i, w_i)$ and follow the above procedure. In both cases, we estimate $\theta$ and $F$ using the Monte Carlo approach described at the beginning of this section.

In, 1997, Bilker and Wang generalized the Efron's method of bootstrapping to a right-censored and left-truncated data using the nonparametric estimate of the joint distribution of the truncation and censoring times and the nonparametric maximum likelihood estimate for the survival curve [11].

Resampling from a real dataset is an example of the nonparametric bootstrap. When independent samples are drawn from a prespecifed distribution, the bootstrap becomes parametric. Although the estimation procedure does not change, the parametric bootstrap is usually more efficient than its nonparametric counterpart [41].

In this thesis, we will use parametric bootstrap techniques for generating a bootstrap lifetime data from the length-biased Weibull and log-normal distributions. We will apply a fixed censoring mechanism mostly for convenience. Since the truncation distribution, in our case, is uniformly distributed and the survival times are assumed to come from a known distribution, we modified the Bilker and Wang method to suit our approach. Detailed algorithm for the simulation of a right-censored and length-biased (right-censored uniformly left truncated) data is given in the next section. To validate our methodological approach and assess an accuracy of the estimates, we will perform bootstrap simulations using a real datasets from the Canadian

Studies of Health and Aging on dementia. All simulations and estimation procedures will be performed using R software.

We start the chapter with a description of how to generate unbiased and length-biased samples, both with and without covariates, from the two distributions used in this thesis: Weibull and log-normal. Then, we will proceed to more involved simulations. First, we will assess the behavior of our two-sample statistic, an adapted version of the Smirnov two-sample test, and establish a critical region for rejecting the hypothesis of distributional equality of survival in the prevalent and incident samples. Once comparisons of distributions of the survival in incident and prevalent data are made, we will look at the behavior of a maximum likelihood function in a combined sample in terms of its accuracy and efficiency. The obtained maximum likelihood estimates will be compared to those from incident and prevalent likelihoods. Finally, performance of combined likelihood function in estimating regression coefficient $\boldsymbol{\beta}$ will be investigated by comparing a bias of the coefficients estimates for different starting values of $\boldsymbol{\beta}$.

The results of the simulation studies will also be presented.

## 6.1 Basic algorithms

### 6.1.1 Simulations from unbiased Weibull and lognormal distributions

Under certain assumptions, simulations from incident population is quite simple. As mentioned earlier, we will assume that survival times $X_i$ in incident sample are mutually independent and also independent of censoring times, $C_i$. Therefore, for an observation $i$, the probability that at given time $x_i$, the subject will experience an event of interest or will be censored is the same. In simulations of unbiased data, we will adopt a random censoring mechanism because of its simplicity, compliance with the real data scenario, and sufficiency for validating our approach. In contrast to fixed censoring, in which each unit has a potential maximum observation time $x_i$ which may differ from one case to the next but is nevertheless fixed in

advance, in random censoring, we let the censoring times follow a specific distribution. The probability that subject $i$ will be censored at the end of his or her observation time is $S(c_i)$, and censoring times are assumed to be uniformly distributed over a given time interval. Therefore, in order to simulate right-censored observations, we first need to simulate a lifetime vector and, independently, a censoring time vector from prespecifed distributions. Then, we observe which comes first, i.e. which value is smaller: if $X_i < C_i$, we say that an event occurred; otherwise, the observation is said to be censored. We have chosen 1000 bootstrap samples for all the simulations based on results of a preliminary testrun in which the differences in estimates for two different values of $B$, $1,000$ and $10,000$, were minimal (less than $10^{-4}$).

In both cases, the number of observations and the proportion of censoring was chosen to mimic the real CSHA datasets. For simulations of the datasets with a binary covariate (gender), the proportion of men and women was also kept the same as in the real dataset. Since we used parametric simulations, the parameters used in the simulations were first estimated from the data. As mention above, to simulate an incident data we can directly use the unbiased forms of the Weibull and log-normal distributions described below. Detailed setting for the simulations are presented later in this section.

**Distributions without covariates**

Let's first look at the unbiased forms of the two distributions, Weibull and log-normal, followed by a description of the algorithm used to simulate bootstrap samples from those distributions.

The unbiased distribution of the Weibull $(\alpha, \lambda)$ has the following form:

$$f_u(x) = \alpha\lambda^\alpha x^{\alpha-1} e^{-(\lambda x)^\alpha}, \quad \alpha, \lambda > 0, x \geq 0. \tag{6.2}$$

The unbiased distribution of the Log-normal $(\mu, \sigma)$ is:

$$f_u(x, \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\log(x)-\mu)^2}{2\sigma^2}}, \quad x \geq 0. \tag{6.3}$$

An algorithm for simulating samples from the above distributions is given below:

**Algorithm 1.** *To obtain a bootstrap sample of unbiased data with right censoring* ***without*** *covariates from a sample of size* $n$*:*

- *Estimate values for $\alpha$ and $\lambda$, and fix $n$ equal the sample size of the original data*
- *Generate $n$ unbiased times $t_i$ from the appropriate distribution (e.g. Weibull)*
- *Generate $n$ censoring times $c_i$ from the uniform distribution*
- *Let $x_i = t_i \wedge c_i$ and set $\delta_i = I(t_i < c_i)$*
- *Proceed to the Algorithm 2*
- *Verify the proportion of censoring in the obtained dataset: if it is approximately equal to the proportion of censored data in the original dataset then STOP; if not, proceed to the next step*
- *Rescale the uniform distribution in the second step*
- *Repeat the procedure until the proportion of censored observations in the simulated data is approximately equal to the proportion of censoring in the real dataset.*

**Algorithm 2.** *To obtain a desired proportion of censoring in the data simulated by Algorithm 1:*

- *Generate $b$ datasets using the first 4 steps of the Algorithm 1*
- *For each dataset, calculate the proportion of censored observations $p_i = \frac{\sum_i^n \delta_i = I(c_i < t_i)}{n}$*
- *Let $x_i = t_i \wedge c_i$ and set $\delta_i = I(t_i < c_i)$*
- *Calculate the mean of the proportions obtained in the previous steps: $\sum_i^b \frac{p_i}{b}$*
- *Go back to Algorithm 1.*

**Distributions with covariates**

Recall that in the presence of covariates, the density function for AFT models can be written in terms of a baseline density function [32]:

$$f_U(x|\mathbf{z};\beta) = f_0(te^{-\boldsymbol{\beta}'\mathbf{z}})e^{-\boldsymbol{\beta}'\mathbf{z}}. \tag{6.4}$$

Therefore, the Weibull $(\alpha, \lambda, \boldsymbol{\beta})$ becomes:

$$f_U(x|\mathbf{z}) = \alpha\lambda^\alpha(xe^{-\boldsymbol{\beta}'\mathbf{z}})^{\alpha-1}e^{-(\lambda xe^{-\boldsymbol{\beta}'\mathbf{z}})^\alpha}e^{-\boldsymbol{\beta}'\mathbf{z}}. \tag{6.5}$$

By writing $f_U(x|\mathbf{z})$ as

$$f_U(x|\mathbf{z}) = \alpha(\lambda e^{-\boldsymbol{\beta}'\mathbf{z}}x)^\alpha x^{\alpha-1}e^{-(\lambda e^{-\boldsymbol{\beta}'\mathbf{z}}x)^\alpha}, \tag{6.6}$$

we can see that in the case of the AFT models, the Weibull distribution with covariates $f_U(x|\mathbf{z})$ is also a Weibull distribution with parameters $\alpha$ and $\lambda = \lambda e^{-\boldsymbol{\beta}}$.

As explained in [9], when covariates are discrete, the incident population comes from a finite set of distinct Weibull distributions with the same shape parameter $\alpha$, but a different rate parameter $\lambda(\mathbf{z})$ depending on a baseline rate $\lambda$ and a covariate effect $\boldsymbol{\beta}$.

For the purpose of simulations, we can look at the joint density function of lifetimes and covariates:

$$f_U(x, \mathbf{z}) = f_U(x|\mathbf{z})p_z(\mathbf{z}), \tag{6.7}$$

where $f_U(x|\mathbf{z})$ is given as above and $p_z(\mathbf{z})$ is the marginal covariate probability, i.e.

$$p_z(z) = p \text{ if } z = 1 \quad \text{and} \quad (1-p) \text{ if } z = 0. \tag{6.8}$$

Therefore, when covariate $z$ is observed, we know exactly to which population it belongs, and from which Weibull distribution it comes from. As a result, we can simulate the covariate first, and depending on its value we can simulate a corresponding lifetime value using the appropriate Weibull distribution [9].

Below, we will show how to simulate data with covariates when the survival times come from the unbiased Weibull population.

**Algorithm 3.** *To obtain a bootstrap sample with right censoring and* ***a binary covariate*** *from a sample of size $n$ from the unbiased Weibull $(\alpha, \lambda, \beta)$ distribution:*

- *Estimate values for $\alpha$, $\lambda$, and $\beta$, and fix $n$ equal to the sample size of the original data*
- *Generate $n$ covariates from Bernoulli distribution with probability of success $p$ estimated from the incident population*
- *For each covariate $\mathbf{z}_i$, generate an unbiased time $t_i$ from the Weibull ($\alpha$, $\lambda e^{-\beta \mathbf{z}}$)*
- *Generate the censoring times $c_i$ from an appropriately scale uniform distribution*
- *Let $x_i = t_i \wedge c_i$, and set $\delta_i = I(t_i < c_i)$.*

The variable $\delta_i$ is an indicator of whether an event (if $\delta_i = 1$) or censoring (if $\delta_i = 0$) occurred. The obtained data will have a form of $(x_i, \mathbf{z}_i, \delta_i)$.

A similar procedure can be used to obtain a bootstrap sample with covariates from a log-normal distribution. Again, the parameters $\mu, \sigma^2$ have to be first estimated from the real data. The remaining procedure follows the Algorithm 3 with the exception of survival times being generated from a log-normal instead of a Weibull distribution.

In the presence of covariates (based on 6.4), the conditional log-normal distribution function utilized in the simulations has the following form:

$$f_u(x|\mathbf{z}) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\log(x)-\boldsymbol{\beta}\mathbf{z}-\mu)^2}{2\sigma^2}}, \quad x \geq 0. \tag{6.9}$$

### 6.1.2 Simulation from length-biased Weilbull and log-normal distributions with and without covariates

For length-biased data, the simulation procedure is the same as described in the previous section with the exception of simulating the data from length-biased Weibull and log-normal distributions rather than from the unbiased distributions directly. Therefore, we will first derive expressions for the two length-biased distributions with and without covariates and then using

Correa & Wolfson (1999) approach for generating length-biased data, we will show how to generate a length-biased sample. We will then follow with a description of the algorithms used to simulate samples from these distributions.

**Length-biased Weibull**

Recall that the length-biased distribution function of a random variable $Y$ is given by

$$F_Y(t) = \int_0^t \frac{x dF_U(x)}{\mu}, \quad t \geq 0, \tag{6.10}$$

where $\mu$ is the overall mean survival time. The length-biased density function can be then written as

$$f_{LB}(x) = \frac{x f_U(x)}{\mu}, \quad t \geq 0. \tag{6.11}$$

Recall further that both the unbiased and length-biased Weibull distributions belong to the family of Generalized Gamma distributions and can be written as $GG(1, \lambda, \alpha)$ and $GG(1 + 1/\alpha, \lambda, \alpha)$, respectively [15]. This property is very useful when generating data from the length-biased Weibull distribution. It utilizes the unbiased $GG$ distribution and a transformation method shown below instead of simulating directly from the length-biased form of the Weibull.

Let $GG(k, \lambda, \alpha)$ be the unbiased distribution function of $X$ given by

$$f_X(x) = \frac{1}{\Gamma(k)} \lambda\alpha(\lambda x)^{k\alpha-1} e^{-(\lambda x)^\alpha}. \tag{6.12}$$

We have already seen that the length-biased Weibull density function with $\mu = \Gamma(1 + 1/\alpha)/\lambda$ has the following form:

$$f_{LB}(x) = \frac{\lambda}{\Gamma(1 + \frac{1}{\alpha})} \alpha\lambda^\alpha x^\alpha e^{-(\lambda x)^\alpha}. \tag{6.13}$$

This density can be rewritten as

$$f_{LB}(t) = \frac{1}{\Gamma(1 + \frac{1}{\alpha})} \alpha\lambda(\lambda x)^{\alpha(1+\frac{1}{\alpha})-1} e^{-(\lambda x)^\alpha}, \tag{6.14}$$

which is exactly the $GG(1 + 1/\alpha, \lambda, \alpha)$ mentioned above.

Now, let $F_{LB}(x)$ be a length-biased Weibull distribution function given by

$$F_{LB}(x) = \int_0^x \frac{1}{\Gamma(1+\frac{1}{\alpha})} \alpha\lambda(\lambda s)^{\alpha(1+\frac{1}{\alpha})-1} e^{-(\lambda s)^\alpha}, \tag{6.15}$$

and let $u = (\lambda s)^\alpha$. Then,

$$F_{LB}(x) = \int_0^{(\lambda x)^\alpha} \frac{1}{\Gamma(1+\frac{1}{\alpha})} \alpha\lambda u^{\alpha(1+\frac{1}{\alpha})-1} e^{-u} du. \tag{6.16}$$

If we let $h(x) = (\lambda x)^\alpha$, $F_{LB}(x) = F_Z(h(x))$ for $x \geq 0$, and $F_Z$ has the distribution function of Gamma$(1 + 1/\alpha, 1)$ [15].

Moreover, let's define the inverse function of $h(x)$ as $g(s) = s^{1/\alpha}/\lambda$. Then, the following relation holds: if $Z$ is a Gamma$(1 + 1/\alpha, 1)$ r.v., then $W = g(Z)$ will have the desired length-biased Weibull distribution as shown below:

$$P(W \leq x) = P(g(Z) \leq x) = P(Z \leq h(z)) = F_Z(h(x)) = F_{LB}(x). \tag{6.17}$$

Therefore, we can simplify the process of generating a length-biased sample from the length-biased Weibull distribution as described by Algorithm 4.

**Algorithm 4.** *To obtain a bootstrap sample of size $n$ with right censoring and left-truncation* ***without*** *covariates from a length-biased Weibull distribution:*

*Let*

$$g(s) = s^{1/\alpha}/\lambda$$

- *Estimate values for $\alpha$, $\lambda$, and fix $n$ equal to the sample size of the original data.*
- *Generate $n$ times $w_i$ from a Gamma$(1 + 1/\alpha, 1)$.*
- *Set $Y = g(W)$ which is the desired length-biased Weibull r.v.*
- *For each $y_i$, generate the truncation time $t_i$ from a $U(0, y_i)$.*
- *Set $r_i = y_i - t_i$, the residual lifetime for each observation.*
- *Pick a constant censoring time $c$ and let $c_i = c$.*
- *Let $x_i = t_i + r_i \wedge c_i$ and set $\delta_i = I(r_i < c_i)$*

### Length-biased Weibull with covariates

In this section, we will discuss how to simulate length-biased data with covariates when lifetime variables come from a mixture of Weibull distributions. As explained for the unbiased case, for a discrete covariate $z$, for example gender, the Weibull distributions have the same shape parameter $\alpha$, but a different rate parameter $\lambda(\mathbf{z})$ that depends on a baseline rate $\lambda$ and coefficients $\boldsymbol{\beta}$. However, in the length-biased case, the setting is slightly different: first, the length-biased mixture of distributions come from $GG$ distributions and second, the marginal distribution of covariates depend not only on a covariate effect $\boldsymbol{\beta}$, but also on the parameters of interest $\alpha$ and $\lambda$. Recall, that a length-biased distribution of covariates is given by

$$f_B(\mathbf{z};\alpha,\lambda,\boldsymbol{\beta}) = \frac{\mu(\mathbf{z};\alpha,\lambda,\boldsymbol{\beta}) f_Z(\mathbf{z})}{\mu(\alpha,\lambda,\boldsymbol{\beta})} = \frac{e^{\beta'\mathbf{z}} f_Z(\mathbf{z})}{E(e^{\beta'\mathbf{z}})}, \tag{6.18}$$

where $f_Z(\mathbf{z})$ is the unbiased distribution of covariates, $\mu(\alpha,\lambda,\boldsymbol{\beta})$ is the overall mean lifetime of the unbiased population, and $\mu(\mathbf{z};\alpha,\lambda,\boldsymbol{\beta})$ is the mean lifetime of an unbiased population depending on the value of covariate $\mathbf{z}$. Note, that if $z$ is a binary covariate, then the distribution of $z$ follows Bernoulli distribution with probability $p$ being estimated from the incident data.

A joint length-biased distribution of lifetimes and covariates is given by the following equation:

$$f_{LB}(x,\mathbf{z};\alpha,\lambda,\boldsymbol{\beta}) = \frac{x f_U(x|\mathbf{z};\alpha,\lambda,\boldsymbol{\beta})}{\mu(\mathbf{z};\alpha,\lambda,\boldsymbol{\beta})} f_B(\mathbf{z};\alpha,\lambda,\boldsymbol{\beta}), \tag{6.19}$$

and the mixture probabilities in this case are just the marginal probabilities of the covariate [9].

Therefore, in order to simulate the desired data, we first estimate $f_B(\mathbf{z})$ based on 6.18 using an appropriate distribution followed by simulation of lifetimes from a length-biased Weibull for each value of $z$.

**Algorithm 5.** *To obtain a bootstrap sample of size n of length-biased data with uniform left truncation and right censoring from Weibull distribution* ***with covariates****:*

*Let*

$$g(s) = s^{1/\alpha}/\lambda$$

- *Estimate values for $\alpha$, $\lambda$, $\boldsymbol{\beta}$ and fix $n$ equal to the sample size of the original data.*

- *Generate n covariates $z$ from 6.18.*
- *For each $\mathbf{z}_i$, generate $n$ lifetimes $w_i$ from an appropriate Gamma$(1 + 1/\alpha, 1)$.*
- *Set $Y = g(W)$ which is the desired length-biased Weibull r.v.*
- *For each $y_i$, generate the truncation time $t_i$ from $U(0, y_i)$.*
- *Set $r_i = y_i - t_i$, the residual lifetime for each observation.*
- *Pick a constant censoring time $c$ and let $c_i = c$.*
- *Let $x_i = t_i + r_i \wedge c_i$ and set $\delta_i = I(r_i < c_i)$*

As before, $\delta_i$ is an indicator of an event $(\delta_i = 1)$ or censoring $(\delta_i = 0)$. The resulting data will have the form of $(x_i, \mathbf{z}_i, \delta_i)$.

**Length-biased log-normal**

As with the Weibull distribution, we will show that a length-biased log-normal r.v. can also be generated using a simpler distribution. In the case of log-normal distribution, we will use the normal distribution, $N(\mu + \sigma^2, \sigma^2)$.

Based on the equations 6.10 and 6.11, a length-biased log-normal distribution function with the overall mean in an unbiased population $\mu = e^{\mu + \frac{\sigma^2}{2}}$, is

$$F_{LB}(x) = \int_0^x \frac{1}{e^{\mu+\frac{\sigma^2}{2}}} \frac{x}{x\sigma\sqrt{2\pi}} e^{\frac{(lnx-\mu)^2}{2\sigma^2}} dx. \tag{6.20}$$

Now, by the change of variables: $y = logx$, the $cdf$ can be written as

$$F_{LB}(y) = \int_0^y \frac{e^y}{\sigma\sqrt{2\pi}} e^{\frac{(y-\mu)^2}{2\sigma^2}} e^{-\mu-\frac{\sigma^2}{2}} dy. \tag{6.21}$$

Simple algebraic manipulation shows that $F_{LB}(y)$ is equivalent to the following:

$$F_{LB}(y) = \int_0^y \frac{1}{\sigma\sqrt{2\pi}} e^{-y^2-2(\mu+\sigma^2)y+(\mu+\sigma^2)^2} dy = \int_0^y \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y-(\mu+\sigma^2))^2}{2\sigma^2}} dy, \tag{6.22}$$

which is the desired normal distribution with mean $\mu = \mu + \sigma^2$ and variance $\sigma^2 = \sigma^2$.

Next, if we let $Z$ to be a r.v. from the above normal distribution, then by the property 6.17, $W = g(Z) = e^Z$ will be the length-biased log-normal r.v.

The simulation procedure from a length-biased log-normal distribution is outlined in Algorithm 5. Only first four steps are presented since the remaining procedure is the same as in Algorithm 4.

**Algorithm 6.** *To obtain a bootstrap sample of size $n$ of length-biased data with right censoring and left-truncation from log-normal distribution* ***without*** *covariates:*

*Let*

$$g(s) = e^s$$

- *Estimate values for $\alpha$, $\lambda$, and fix $n$ equal the sample size of the original data.*
- *Generate $n$ times $w_i$ from a Normal$(\mu + \sigma^2, \sigma^2)$.*
- *Set $Y = g(W)$ which is the desired length-biased log-normal r.v.*
- *Proceed as in Algorithm 4.*

## 6.2 Validating assumption of distributional equality of incident and prevalent populations

In a perfect sampling scenario, prevalent and incident samples would represent the same underlying population, and inferences drawn based on one population could be applied to the other one and vice versa. However, due to different factors such as sampling strategy, methods of estimation, or existing differences between the two populations, we expect that the estimated survival in the two samples will differ. The main goal of these simulations is to validate our approach of combining the prevalent and incident samples. However, since a formal test to compare two distributions functions in the presence of censoring was not yet developed, we adapt the Smirnov test to our scenario and perform simulations to obtain a distribution of

the statistic under the null hypothesis (i.e. when prevalent and incident data come from the same distribution), and in consequence, determine a rejection region for our "ad hoc" two-sample Smirnov test. In other words, we have to find a value of the $D$ - statistic, $d_{critical}$, such that the probability of observing a distance greater than $d_{critical}$ is less than $0.05$, i.e., $p = \frac{\sum_i^{1000}(d_{\max_i} > d_{critical})}{1000} > 0.05$.

**Methods** The two populations of interest, in our case, are the prevalent (length-biased) and the incident (unbiased) populations. Particularly, we will use the test to compare two non-parametric estimators of the survival functions in these populations: the NPMLE estimator for the length-biased data and the Kaplan-Meier estimator for the unbiased data. Then, we will compare the observed maximum distance with the frequency distribution of maximum distances obtained from the simulation study. We will assume that if two samples come from the same underlying population (the null hypothesis of the test), then the distances between their distribution functions (and consequently between their survival functions) are relatively small.

Since Weibull distribution was previously proved to fit the CSHA prevalent population well, we will use it as a reference distribution for our simulations [48]. We will therefore assume that incident and prevalent data come from the same Weibull distribution with the shape and scale parameters, $\alpha = 1.24$ and $\lambda = 0.19$, estimated from the real data using maximum likelihood approach. Also, since the simulations are performed to investigate the behavior of the statistic in a particular scenario, the size of the bootstrap samples and the proportion of censored observations will also be chosen to resemble the real datasets. To perform simulations, we will generate $B = 1000$ bootstrap samples of each, the incident $(n = 480, p_{event} = 52\%)$ and prevalent $(m = 896, p_{event} = 78\%)$ datasets from the unbiased and length-biased Weibull distributions, respectively using Algorithm 1 and Algorithm 4 described in the previous section. Then, we will calculate the maximum distances between the NPMLE developed by Vardi and the KM estimator for each pair of the simulated incident and the prevalent samples (1000 in total). Finally, we we will use the frequency distribution of the obtained $1000$ maximum distances ($D$-statistic) to assess the probability of observing the distance we have seen in the real

data. The proportion of maximum distances greater than the distance observed in the real data will serve as an approximation to the p-value for either rejecting or accepting the hypothesis of equality of survival distributions in the prevalent and incident populations. In other words, if based on the appropriate critical value, the observed distance is significantly different from the maximum distance obtained from the two samples that came from the same distribution, then we would reject the null hypothesis of the test.

A detailed algorithm used in above simulations is presented below.

**Algorithm 7.** *To obtain a critical region for a two-sample K-S type of test based on $B$ bootstrap samples from each unbiased and length-biased Weibull distributions:*

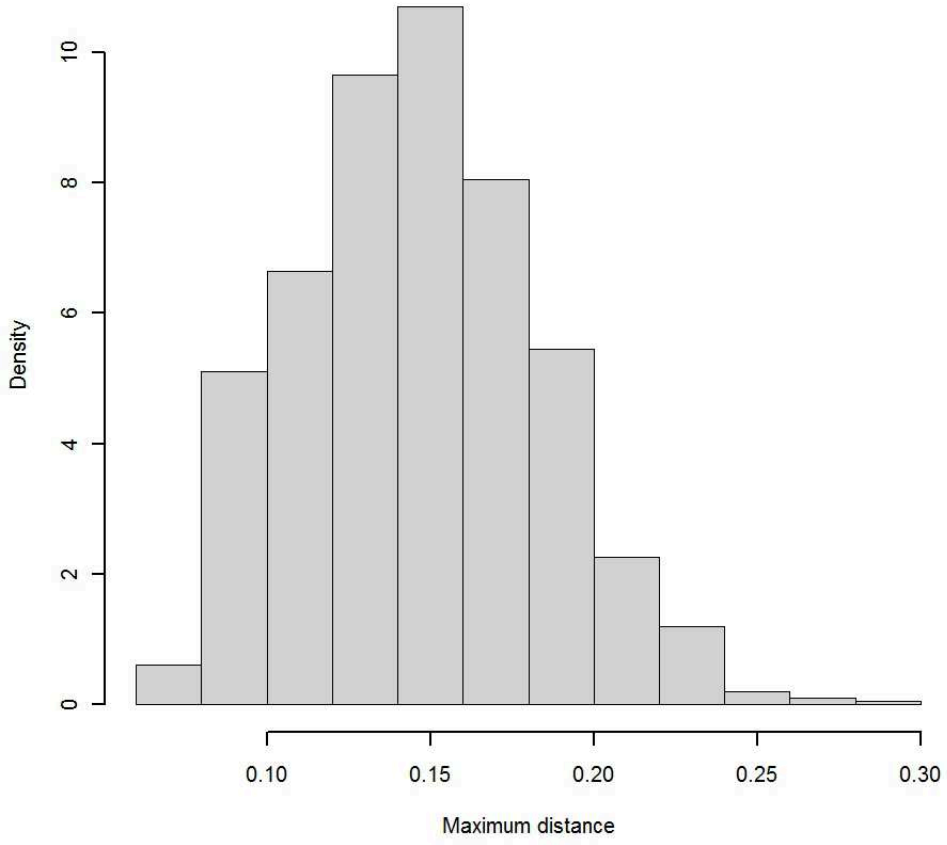
- *Estimate values for $\alpha$ and $\lambda$ from the appropriate data, and set $n$ and $m$ to be equal to the sizes of a real incident and a prevalent datasets, respectively.*
- *Generate $n$ bootstrap samples from an unbiased Weibull distribution using Algorithm 1.*
- *Generate $m$ bootstrap samples from a length-biased Weibull distribution using Algorithm 4.*
- *Extract vectors of lifetimes $t_I = (t_{I_1} \dots t_{I_n})$ and survival times $S_I = (S_{I_1} \dots S_{I_n})$ from the unbiased data, and vectors of lifetimes $(t_P = t_{P_1} \dots t_{P_m})$ and survival times $(S_P = S_{P_1} \dots S_{P_m})$ from the length-biased data.*
- *Define step function for each dataset as $f_I(t_I) = S_I(i)$ for $t_I$ in $(t_{I_i}, t_{I_{i+1}})$ and $f_P(t_P) = S_P(j)$ for $t_P$ in $(t_{P_j}, t_{P_{j+1}})$, respectively.*
- *Define a common lifetimes supervector $t_{All} = t_I \cup t_P$.*
- *For each $t_{ALL_k}$ calculate a distance $d_k = f_I(t_I) - f_P(t_P)$.*
- *Find the maximum distance, $d_{\max} = \max(d_k)$.*
- *Repeat the above steps $B$ times.*

- *Set p-value as the proportion of maximum distances bigger than the maximum distance calculated from the real dataset, i.e.*

$$p = \frac{\sum_i^B (d_{\max_i} > d_{real})}{B}$$

**Results** A histogram of maximum distances obtained from the simulations is presented in Figure 6.1.

Figure 6.1: Histogram of maximum distances between incident and prevalent samples from 1000 simulations



Based on the distribution of maximum distances in the above simulations, we would reject the null hypothesis of distributional equality of incident and prevalent populations if the maximum distance between the incident and prevalent survival functions was greater than $0.21$ (critical value for $95\%$ level of confidence).

Although the survival functions in the prevalent and incident samples are not statistically different, the distances between the survival functions are quite large despite the large sample size (N=1,375) and the fact that they are drawn from the same distribution. This might be due to the fact that censoring in the incident population is quite heavy (50%) and the Kaplan-Meier estimator do not perform as well as it would in the case of lighter censoring. For later time intervals, when most subjects have either died or been censored, the number of subjects that survived to the beginning of that interval (risk set) is small, and the variance estimates for these intervals are generally less reliable than those for earlier intervals i.e. they will tend to underestimate the actual variance. In fact, the maximum distances observed in the simulations decreased substantially as the number of events in either population was increasing. Also, the NPMLE estimator proposed by Vardi adjusts for the length bias by weighting the times inversely and putting more weight on early survival times, which, in return, may induce bigger distances between the incident and prevalent populations for early time intervals. Taking into account that the largest distances between the survival curves including the maximum distance of 0.337 occur for very small survival times (approximately 6 months), our suggestion that the instability of the Vardi's NPMLE for small survival times, could inflate the real distances between the curves seems to be correct. It is also possible that other factors, such as gender or age, have an impact on the distribution of observed distances. When stratified by gender, we noticed that the distances were much smaller for women than for men. However, the women sample size was approximately twice as big as that of men ($n_F = 324$ vs $n_M = 156$ in the incident case and $n_F = 630$ vs $n_M = 266$ in the prevalent case).

To summarize, under the specific scenario, our two-sample test fail to reject a null hypothesis at $\alpha = 0.05$ level of confidence. However, it would be helpful to extend the simulations to scenarios that include different number of censoring in both incident and prevalent data to

observe how the amount of censored observations in a sample impacts the distribution of maximum distances, and consequently, the cut off point for the statistic. Assessing statistical power of the test would help to determine how good the test performs in a given scenario. Further studies incorporating some of the potential predictors of the survival and sampling weights would also be desired.

## 6.3 Assessing performance of a combined likelihood function

Once we establish that the incident and prevalent samples represent the same underlying population, we are able to use a combined likelihood function in estimation of relevant parameters and regression coefficients. However, in order to draw valid conclusions based on the MLEs from the joint likelihood function, we should make sure that the discrepancy between the estimates from incident, prevalent, and combined datasets can be attributed to the real differences that exist between these populations rather than to the approach we undertook to obtain the estimates. In addition, since no asymptotic properties for a combined likelihood function were developed to this time, and even if they were developed, the behavior of the likelihood in a case of finite sample would still be unknown, we use simulations to address this issue. The interest here is to evaluate relative efficiency of $L_{Comb}$ to $L_{Prev}$ and $L_{Inc}$ in terms of the ratio of the mean square errors (MSE) of the estimates.

**Methods** We generate $1000$ replicates of each incident, prevalent, and combined datasets with the size and proportion of censoring equivalent to the values observed in the CSHA samples. To generate the three datasets, we will use two sets of parameters, $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$, corresponding to the Weibull ($\alpha = 1.24, \lambda = 0.19$) and Weibull ($\alpha = 2.6, \lambda = 0.19$), respectively. These parameters are equivalent to the parameters estimated from the prevalent and incident CSHA data. To improve estimation of the survival, we let the survival functions depend on a binary covariate. The effect of gender, the parameters $\beta_{Prev} = -0.25$ and $\beta_{Inc} = -0.17$, are

estimated from the CSHA data using a joint likelihood of covariates and lifetimes.

The algorithm to generate the required samples is as follows: we first generate a vector of the covariate using the Bernoulli distribution with a probability of success $p = 0.03$ based on a distribution of the covariate in the incident data. For each value of the covariate, we then generated survival times, censoring times (based on the random or fixed censoring mechanism), and additionally uniform truncation times in a case of the prevalent sample. To generate the unbiased and length-biased data we use algorithms described in the previous sections. For the combined dataset, we simply merge already generated incident and prevalent samples into one sample and add the indicator variable, $\xi_i$ with values $\xi_i = 1$ if an observation $i$ cames from the incident sample, and $\xi_i = 0$, otherwise. For each type of data, we used respectively unbiased (incident data), length-biased (prevalent data), and combined (combined incident and prevalent data) maximum likelihood approach to estimate the quantities of interest. Maximum likelihood functions for the Weibull and log-normal distributions used in the simulations are presented in the 6.1.

Table 6.1: Likelihood functions for unbiased, length-biased, and combined data in the presence of a binary covariate

| | *Unbiased* $\mathcal{L}(\boldsymbol{\theta})$ |
|---|---|
| Weibull | $(\alpha\lambda^{\alpha}x^{\alpha-1}e^{-\boldsymbol{\beta}\mathbf{z}\alpha})^{\delta}e^{-(\lambda xe^{-\boldsymbol{\beta}\mathbf{z}})^{\alpha}}$ |
| Log-normal | $\left(\frac{1}{x\sqrt{2\pi}\sigma}e^{-0.5(\frac{\log x-\boldsymbol{\beta}'\mathbf{z}-\mu}{\sigma})^2}\right)^{\delta}\left(1-\Phi\left(\frac{\log x-\boldsymbol{\beta}'\mathbf{z}-\mu}{\sigma}\right)\right)^{1-\delta}$ |

| | *Length-biased joint* $\mathcal{L}(\boldsymbol{\theta})$ |
|---|---|
| Weibull | $\frac{1}{\mu(\boldsymbol{\theta})}(\alpha\lambda^{\alpha}x^{\alpha-1}e^{-\boldsymbol{\beta}\mathbf{z}\alpha})^{\delta}e^{-(\lambda xe^{-\boldsymbol{\beta}\mathbf{z}})^{\alpha}}\left(\frac{p}{1-p}\right)^{z_1}(1-p)$ |
| Log-normal | $\frac{1}{\mu(\boldsymbol{\theta})}\left(\frac{1}{x\sqrt{2\pi}\sigma}e^{-0.5(\frac{\log x-\boldsymbol{\beta}'\mathbf{z}-\mu}{\sigma})^2}\right)^{\delta}\left(1-\Phi\left(\frac{\log x-\boldsymbol{\beta}'\mathbf{z}-\mu}{\sigma}\right)\right)^{1-\delta}\left(\frac{p}{1-p}\right)^{z_1}(1-p)$ |

| | *Combined* $\mathcal{L}(\boldsymbol{\theta})$ |
|---|---|
| Weibull | $\left[\frac{1}{\mu(\boldsymbol{\theta})}\right]^{1-\xi}(\alpha\lambda^{\alpha}x^{\alpha-1}e^{-\boldsymbol{\beta}\mathbf{z}\alpha})^{\delta}e^{-(\lambda xe^{-\boldsymbol{\beta}\mathbf{z}})^{\alpha}}\left(\frac{p}{1-p}\right)^{z_1}(1-p)$ |
| Log-normal | $\left[\frac{1}{\mu(\boldsymbol{\theta})}\right]^{1-\xi}\left(\frac{1}{x\sqrt{2\pi}\sigma}e^{-0.5(\frac{\log x-\boldsymbol{\beta}'\mathbf{z}-\mu}{\sigma})^2}\right)^{\delta}\left(1-\Phi\left(\frac{\log x-\boldsymbol{\beta}'\mathbf{z}-\mu}{\sigma}\right)\right)^{1-\delta}\left(\frac{p}{1-p}\right)^{z_1}(1-p)$ |

$\mu(\boldsymbol{\theta})$ for the length-biased Weibull and log-normal distributions are given by the following:

$$\mu(\boldsymbol{\theta})_{Weib}=\int_{\mathbf{z}}\mu(\mathbf{z};\boldsymbol{\theta})f_z(\mathbf{z})dz=\frac{\Gamma(1+\frac{1}{\alpha})}{\lambda}\left[pe^{\beta_1}+(1-p)\right], \tag{6.23}$$

and

$$\mu(\boldsymbol{\theta})_{Log}=\int_{z}\mu(\mathbf{z};\boldsymbol{\theta})f_z(\mathbf{z})dz=e^{\mu+0.5\sigma^2}\left[pe^{\beta_1}+(1-p)\right]. \tag{6.24}$$

The detailed calculations of $\mu(\boldsymbol{\theta})$ are postponed to the Application chapter where we consider two covariates simultaneously.

To asses a relative efficiency of the combined likelihood, we first compared the MLEs obtained from the simulations of each type of the data to the real parameters values (the values

we simulated the data from), in terms of their MSEs. Next, we used the ratio of the MSE of a given parameter in the $L_{Comb}$ to the MSEs of the same parameters in the $L_{Prev}$ and $L_{Inc}$ to obtain the relative efficiency of the combined likelihood function, and compare its performance to the performance of the incident and prevalent likelihoods separately. For $\boldsymbol{\beta}$ coefficients we also compared their values with the MLEs obtained from the likelihood function for covariates only.

**Results** The relative efficiencies of $\mathcal{L}_{Comb}$ and $\mathcal{L}_{Prev}$ for the initial set of the true parameters $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ are presented in Table 6.2. The relative efficiencies of $\mathcal{L}_{Comb}$ and $\mathcal{L}_{Inc}$ for the same set of the true parameters are presented in Table 6.3. $\hat{\boldsymbol{\theta}}_C$, $\hat{\boldsymbol{\theta}}_P$, and $\hat{\boldsymbol{\theta}}_I$ denote estimates of the parameters' vector in the combined, prevalent, and incident case, respectively.

Table 6.2: Efficiency results of $\mathcal{L}_{Comb}$ vs $\mathcal{L}_{Prev}$ for the initial parameters $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$

| $\boldsymbol{\theta}$ | *Average* $\hat{\boldsymbol{\theta}}_C$ | *Average* $\hat{\boldsymbol{\theta}}_P$ | *Var* $(\hat{\boldsymbol{\theta}}_C)$ | *Var* $(\hat{\boldsymbol{\theta}}_P)$ | *Eff.* $(\hat{\boldsymbol{\theta}}_C : \hat{\boldsymbol{\theta}}_P)$ |
|---|---|---|---|---|---|
| $\alpha_1$ | 1.2414 | 1.2438 | 0.00098 | 0.00222 | 2.2665 |
| $\lambda_1$ | 0.1900 | 0.1900 | 0.00003 | 0.00007 | 2.1877 |
| $\beta_1$ | -0.2504 | -0.2488 | 0.00179 | 0.00251 | 1.4023 |
| $p$ | 0.2998 | 0.2988 | 0.00021 | 0.00038 | 1.8162 |
| $\alpha_2$ | 2.6006 | 2.6044 | 0.00460 | 0.00726 | 1.5827 |
| $\lambda_2$ | 0.1901 | 0.1900 | 0.00001 | 0.00001 | 1.5692 |
| $\beta_2$ | -0.1690 | -0.1688 | 0.00051 | 0.00070 | 1.3812 |
| $p$ | 0.3004 | 0.3004 | 0.00016 | 0.00028 | 1.7278 |

Table 6.3: Efficiency results of $\mathcal{L}_{Comb}$ vs $\mathcal{L}_{Inc}$ for the initial parameters $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$

| $\boldsymbol{\theta}$ | *Average* $\hat{\boldsymbol{\theta}}_C$ | *Average* $\hat{\boldsymbol{\theta}}_I$ | *Var* $(\hat{\boldsymbol{\theta}}_C)$ | *Var* $(\hat{\boldsymbol{\theta}}_I)$ | *Eff.* $(\hat{\boldsymbol{\theta}}_C : \hat{\boldsymbol{\theta}}_I)$ |
|---|---|---|---|---|---|
| $\alpha_1$ | 1.2414 | 1.2427 | 0.00098 | 0.00406 | 4.1288 |
| $\lambda_1$ | 0.1900 | 0.1900 | 0.00003 | 0.00015 | 4.5935 |
| $\beta_1$ | -0.2504 | -0.2519 | 0.00179 | 0.01165 | 6.5146 |
| $p$ | 0.2998 | 0.3010 | 0.00021 | 0.00046 | 2.1997 |
| $\alpha_2$ | 2.6006 | 2.6095 | 0.00460 | 0.01640 | 3.5850 |
| $\lambda_2$ | 0.1901 | 0.1901 | 0.00001 | 0.00003 | 4.2810 |
| $\beta_2$ | -0.1690 | -0.1716 | 0.00051 | 0.00278 | 5.4658 |
| $p$ | 0.3004 | 0.2987 | 0.00016 | 0.00042 | 2.5430 |

As shown in the tables, in the case when data come from the same distribution, the parameters' estimates from the unbiased, length-biased, and combined likelihoods are very close to the true values of $\boldsymbol{\theta}$, i.e. the ones we simulated the data from. The bias of the parameters $\alpha, \lambda$, the regression coefficient $\beta$, and proportion $p$ is less than $10^{-2}$ in all of the simulations. The MLEs from the combined likelihood exhibit smaller biases than the same estimates from the incident and prevalent likelihood functions, separately. The parameters are rather over- then underestimated with the exception for the regression parameter $\beta$ which is slightly underestimated. The overestimation of the parameters is possibly due to the larger number of prevalent cases in the combined data, and the fact that individuals in the prevalent cohort tend to live longer.

Efficiency and precision of the estimates, proved also to be better in the combined likelihood. Relative efficiency of $\mathcal{L}_{Comb}$ is approximately two times higher when compared to $\mathcal{L}_{Inc}$ than to $\mathcal{L}_{Prev}$. This increase might be a result of increased sample size in the case of combined data, as well as decreased variability in the estimates due to a lower proportion of censored observations in the combined vs incident sample. Remember, that the proportion of censored observations in the incident sample is as high as $50\%$, and in the prevalent case, $30\%$. The estimates from the $\mathcal{L}_{Comb}$ are closer to the ones from the $\mathcal{L}_{Prev}$ indicating that $\mathcal{L}_{Comb}$ puts more

weight on the prevalent cases, again possibly due to the larger number of prevalent than incident cases in the combined dataset (895 vs 480). The largest increase in efficiency of $\mathcal{L}_{Comb}$ was observed for the regression parameter $\beta$: four times higher compared to the $\mathcal{L}_{Prev}$, and six times higher compared to the $\mathcal{L}_{Inc}$. It is possible that the size of the effect of a covariate has an impact on the efficiency of the combined likelihood in estimating the parameter. Although the increased sample size could play a main role in improving the efficiency of the estimates, the use of combined likelihood is justified as it leads to more accurate and more efficient estimates.

For each set of parameters, a survival curve from the corresponding Weibull distribution was computed. Figures 6.2 and 6.3 give the relative efficiency of the combined likelihood over the incident and prevalent likelihoods for the survival function at each sample time point. Both figures show a gradual increase in efficiencies of combined likelihood over the prevalent and incident approaches making it overall more efficient as suggested by the bootstrap parameters.

Figure 6.2: Estimated efficiency of $\mathcal{L}_{Comb}$ over $\mathcal{L}_{Prev}$

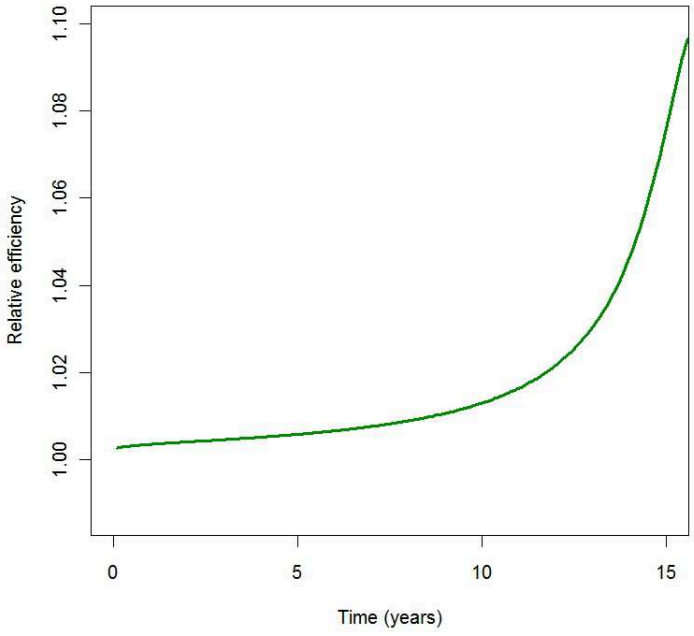
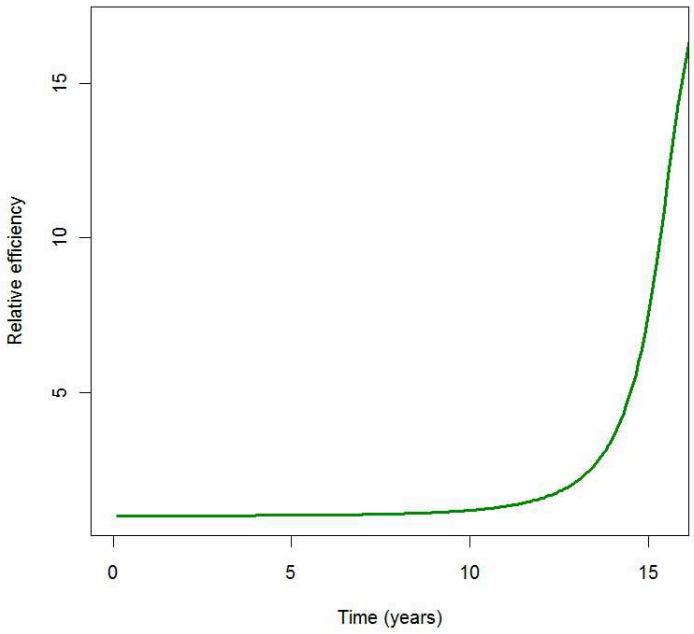


Figure 6.3: Estimated efficiency of $\mathcal{L}_{Comb}$ over $\mathcal{L}_{Inc}$




## 6.4 Assessing the behavior of regression coefficients $\beta$s in the combined likelihood

The performance of a combined likelihood function (and any likelihood based approach for this matter) in estimating regression coefficients can be examined by looking at the behavior of the function when different values of $\boldsymbol{\beta}$ are set to be true. This is specially important for the length-biased part of the combined likelihood function because, as mentioned before, length-biased sampling introduces bias also to a distribution of covariates. For example, if a subpopulation with covariate $Z = 1$ has a longer survival times compared to the one with $Z = 0$, and consequently, a higher probability of being included in the sample, then the effect of this covariate on survival will be most likely overestimated. We would like to see how the size of real $\boldsymbol{\beta}$s effects the accuracy of a combined likelihood approach. The standard deviation of obtained estimates of interest $SE(\hat{\boldsymbol{\beta}})$, can be used to assess uncertainty in the estimates. A compari-

son of the simulated results with the true values of $\boldsymbol{\beta}$ chosen initially provides a measure of the performance (bias) of the simulation process. If the amount of bias vary from $\frac{1}{2}SE(\hat{\beta})$ to $2SE(\hat{\beta})$, it is considered substantial and the performance of the method unsatisfactory [12].

**Methods** In these simulations, we used the combined likelihood function based on a joint distribution of lifetimes $X$ and a covariate $Z$ given by 4.45. We generated $B = 1000$ samples from Weibull ($\alpha = 1.24, \lambda = 0.19$) for five different values of $\boldsymbol{\beta}$: $\{-3, -1, 0, 1, 3\}$. The number of observations and proportion of censored observations was chosen based on the real dataset ($N = 1375, p_{censored} = 0.31$). For simplicity reason, we have chosen a one-dimensional vector of covariates but the method can be extended to include more than one covariate. Here, a binary covariate $Z$ is set to come from a Bernoulli distribution with probability $p = 0.30$ also estimated from the observed data. Therefore, the distribution of covariates is assumed to be known (Bernoulli (0.3)), and the data come from the population consisting of two distinct subpopulations identified by a binary covariate $Z$ that assumes either value of 1 or 0 with probability in the unbiased population.

Table 6.4 summarizes the results of the simulations.

Table 6.4: Bootstrap parameters estimates from $\mathcal{L}_{Comb}$ for different values of $\boldsymbol{\beta}$s

| *Value of* $\beta$ | *Estimate* | *Variance* | *Bias* |
|---|---|---|---|
| -3 | -3.1203 | 0.0208 | -0.1203 |
| -1 | -1.1111 | 0.0040 | -0.1111 |
| *-0.25* | *-0.2953* | *0.0032* | *-0.0453* |
| 0 | -0.0009 | 0.0028 | -0.0009 |
| 1 | 1.2368 | 0.0032 | 0.2368 |
| 3 | 3.4600 | 0.0060 | 0.4600 |

For a discrete covariate with a known unbiased distribution, the MLEs obtained from $\mathcal{L}_{Comb}$ seem to induce a slight bias. The estimates underestimate $\beta$ for the negative initial values of $\beta$ and overestimate for the positive ones. Variability of the estimates seems to increase

faster and be larger for the negative than for the positive true values of $\beta$. For larger values of $\beta$, the probability of survival for a subpopulation identified with the covariate ($Z = 1$) becomes smaller than in the subpopulation identified by $Z = 0$, and the bias induced by the covariate increases (the subpopulation will be highly underrepresented in the sample). The observed bias is proportional to $|\beta|$, probably because as $|\beta|$ increases, the sampling proportion of the subpopulation with shorter lifetimes goes to zero. It is possible that the combination of censoring of the longer lifetimes and the inverse length bias transformation giving more weight to the shorter-lived subpopulation results in a tendency to overestimate $P_Z(Z = 1)$ in the setting of these simulations [9].

# Chapter 7

# 7. Applications: CSHA study of dementia

## 7.1 The data

The data used for this thesis came from the multicenter epidemiological study of dementia and other health problems in elderly people in Canada [18]. The Canadian Study of Health and Aging (CSHA) was undertaken in three main phases: the CSHA-1 field study in 1991-92, the CSHA-2 five years follow-up study in 1996-97, and the CSHA-3 follow-up in 2001-02. Among the many aims of the study was the estimation of prevalence (phase 1), incidence (phase 2), and risk factors of dementia. In phase 1 of the study, a representative sample of people aged 65 and over was chosen randomly from communities (9,008) and from institutions (1,225). To assess their eligibility for the study, subjects underwent screening test for cognitive impairment (community) and standardized clinical examination (both groups). DSM-III-R and ICD-10 criteria were used to classify patients into three groups: 1) cognitive loss, no dementia (772), 2) normal, no cognitive loss (523), and 3) dementia (1,132). To assess a degree of impairment, further test were implemented for patients diagnosed with dementia. The National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Associations criteria were used to assess probable (448) and possible (301) Alzheimer's categories, and ICD-10 criteria were used to define subcategories of vascular and other dementias [18].

The second phase of the study was conducted in 1996. All the subjects from the first phase that were alive, underwent similar to the phase I evaluation. For those who died between the phases of the study, the data on survival were obtained from their families. Second phase included originally 1,132 subjects in total. Patients with missing date of onset, those with other or unclassified dementia, and two outliers (patients who died more than 50 years after the onset of symptoms) were excluded from the analysis. Ultimately, the prevalent study included 896 patients at the end of Phase 2. The incident sample consisted of healthy subjects (without dementia) who were recruited and followed until the onset of dementia, and then continued to be followed until death or censoring occurred. In total, 480 subjects were identified as incident cases.

## 7.2 Methods and analysis

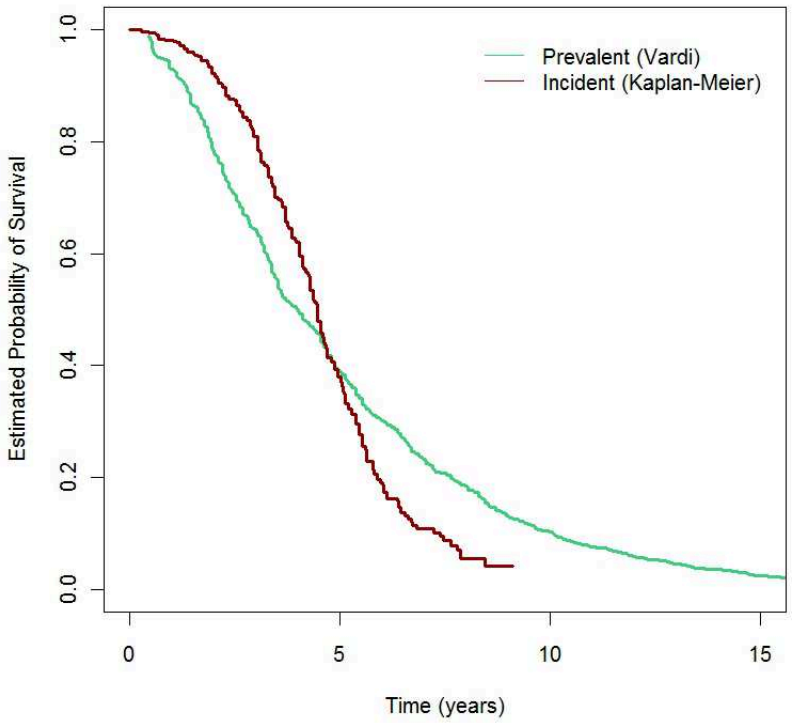
For the purpose of our analysis, we used the prevalent and incident CSHA datasets. Only subjects diagnosed with vascular dementia, and probable and possible Alzheimer's disease were included in the sample. We retained the following variables for our study: approximate date of onset, date of death or censoring, the indicator whether the observation was censored or not, and gender and age at onset (AAO) covariates. The resulted datasets included 896 and 480 subjects for the prevalent and incident study, respectively. Approximately, 78 % of prevalent cases and 51% of incident cases in our samples experienced the event of interest (death from dementia).

The survival times in the incident sample are right-censored variables, and the survival function, in this case, can be estimated using standard survival methods such as the Kaplan-Meier estimator. The situation becomes more complicated in the case of the prevalent data, in which subjects who already experienced dementia were recruited into the study. In this case, the survival times are subjected not only to right censoring (subjects are followed for the prespecifed study period or may be lost to follow up), but also to left truncation (the onset times had occurred before the study begun). Since no particular event that would lead to an outbreak

of dementia was observed, we could reasonably assume that the incident rate of dementia has remained constant and the time from the onset to the beginning of the study are uniformly distributed. Also, since the stationarity assumption that follows from the uniform distribution of truncation times was previously assessed for the CSHA prevalent sample by other authors (see [3]), we readily applied the result in our study.

The survival functions for the incident and prevalent cases were obtained using the Kaplan-Meier estimator and the NPMLE developed by Vardi, respectively (Figure 7.1).

Figure 7.1: Incident and prevalent survival functions in the CSHA sample



As shown in Figure 7.3, the survival estimated based on the KM estimator does not go beyond the last event observed and puts more weight on the tail of the distribution. Since there is a high proportion of censored observations in the incident sample (approximately, 52 %), which in this case is rather due to having the study end while many subjects are still alive than loss to follow-up, the equivalent number of subjects exposed (at risk) at later times decreases, making the Kaplan-Meier estimate less reliable than it would be for the same number of subjects with less censoring. As a result, the estimated variances become poorer approximations, perhaps considerably smaller than the actual variances. In contrast to the KM estimator, the NPMLE estimator (for prevalent cases), which adjusts for the length-biased in a sampling process by putting more weight on the early survival times and therefore tends to underestimate the survival for these times, performs better for later lifetimes. To examine if the observed differences in survival between the incident and prevalent populations can be attributed rather to the performance of the estimators or to existing real differences in characteristics of both populations, than to the fact that the two samples do not represent the same underlying population, we used a two-sample Smirnov type test. The simulation study described in the previous chapter gave us an idea of how large the maximum distance between both curves must be in order to not reject the null hypothesis. We obtained a maximum distance of $d_{\max} = 0.177$ (p-value: $0.20$) from the real datasets, and based on the simulations study, we did not reject the hypothesis of distributional equivalence between the incident and prevalent CSHA samples.

Once we showed that the incident and prevalent samples in the CHSA study have the same underlying distribution, we could combine the two samples into one dataset to improve the efficiency and reliability of survival estimates by utilizing the information in both types of the data simultaneously. We used two parametric distributions, Weibull and log-normal, to obtain a combined MLE of the survival function and compare it to the MLEs obtained from the incident and prevalent data alone. The likelihood functions for the lifetime $X$ coming from the Weibull and log-normal distributions and two covariates, gender and age at onset, can be found in Table 7.1.

Table 7.1: Weibull and log-normal likelihood functions with two covariates (gender, AAO) for the prevalent, incident, and combined CSHA samples and a likelihood function for covariates only

| $X \sim Weibull(\alpha, \lambda)$, $Z_1 \sim Bernoulli(p)$, $Z_2 \sim N(\mu_M, \sigma_M)$, $Z_3 \sim N(\mu_F, \sigma_F)$ |
|---|
| $\mathcal{L}_{Inc}:$ $(\alpha\lambda^{\alpha}x^{\alpha-1}e^{-\boldsymbol{\beta}\mathbf{z}\alpha})^{\delta}e^{-(\lambda xe^{-\boldsymbol{\beta}\mathbf{z}})^{\alpha}}$ |
| $\mathcal{L}_{Prev}:$ $\frac{1}{\mu(\boldsymbol{\theta})}(\alpha\lambda^{\alpha}x^{\alpha-1}e^{-\boldsymbol{\beta}\mathbf{z}\alpha})^{\delta}e^{-(\lambda xe^{-\boldsymbol{\beta}\mathbf{z}})^{\alpha}}\left(\frac{p}{1-p}\right)^{z_1}(1-p)\Phi^{z_1}(z_2)\Phi^{1-z_1}(z_3)$ |
| $\mathcal{L}_{Comb}:$ $\left[\frac{1}{\mu(\boldsymbol{\theta})}\right]^{1-\xi}(\alpha\lambda^{\alpha}x^{\alpha-1}e^{-\boldsymbol{\beta}\mathbf{z}\alpha})^{\delta}e^{-(\lambda xe^{-\boldsymbol{\beta}\mathbf{z}})^{\alpha}}\left(\frac{p}{1-p}\right)^{z_1}(1-p)\Phi^{z_1}(z_2)\Phi^{1-z_1}(z_3)$ |
| $X \sim Log-normal(\mu, \sigma)$, $Z_1 \sim Bernoulli(p)$, $Z_2 \sim N(\mu_M, \sigma_M)$, $Z_3 \sim N(\mu_F, \sigma_F)$ |
| $\mathcal{L}_{Inc}:$ $\left[\frac{1}{x\sqrt{2\Pi}\sigma}e^{-0.5\left(\frac{lnx-\boldsymbol{\beta}\mathbf{z}-\mu}{\sigma}\right)^2}\right]^{\delta}\left[1-\Phi\left(\frac{lnx-\boldsymbol{\beta}\mathbf{z}-\mu}{\sigma}\right)\right]^{1-\delta}$ |
| $\mathcal{L}_{Prev}:$ $\frac{1}{\mu(\boldsymbol{\theta})}\left[\frac{1}{x\sqrt{2\Pi}\sigma}e^{-0.5\left(\frac{lnx-\boldsymbol{\beta}\mathbf{z}-\mu}{\sigma}\right)^2}\right]^{\delta}\left[1-\Phi\left(\frac{lnx-\boldsymbol{\beta}\mathbf{z}-\mu}{\sigma}\right)\right]^{1-\delta}\times$ $\left(\frac{p}{1-p}\right)^{z_1}(1-p)\Phi^{z_1}(z_2)\Phi^{1-z_1}(z_3)$ |
| $\mathcal{L}_{Comb}:$ $\left[\frac{1}{\mu(\boldsymbol{\theta})}\right]^{1-\xi}\left[\frac{1}{x\sqrt{2\Pi}\sigma}e^{-0.5\left(\frac{lnx-\boldsymbol{\beta}\mathbf{z}-\mu}{\sigma}\right)^2}\right]^{\delta}\left[1-\Phi\left(\frac{lnx-\boldsymbol{\beta}\mathbf{z}-\mu}{\sigma}\right)\right]^{1-\delta}\times$ $\left(\frac{p}{1-p}\right)^{z_1}(1-p)\Phi^{z_1}(z_2)\Phi^{1-z_1}(z_3)$ |
| $Z_1 \sim Bernoulli(p)$, $Z_2 \sim N(\mu_M, \sigma_M)$, $Z_3 \sim N(\mu_F, \sigma_F)$ |
| $\mathcal{L}_{Cov}:$ $\left(\frac{e^{\boldsymbol{\beta}\mathbf{z}}}{pe^{\beta_1 sex+\beta_2\mu_M+0.5\sigma_M^2\beta_2^2}+(1-p)e^{\beta_3\mu_F+0.5\sigma_F^2\beta_3^2}}\right)^{1-\xi}\times$ $\left(\frac{p}{1-p}\right)^{z_1}(1-p)\Phi^{z_1}(z_2)\Phi^{1-z_1}(z_3)$ |

The unbiased joint distribution of gender ($Z_1$) and age at onset ($Z_2$) utilized in the above likelihoods can be written as follows:

$$g(z_1, z_2) = \left(\frac{p}{1-p}\right)^{z_1} (1-p)\Phi^{z_1}_{\mu_M,\sigma_M}(z_2)\Phi^{1-z_1}_{\mu_F,\sigma_F}(z_2), \tag{7.1}$$

where $p$ is a probability of being man ($Z_1 = 1$) in the Bernoulli distribution, and $\mu_M, \sigma_M$ and $\mu_F, \sigma_F$ are the means and standard deviations of the normally distributed age at onset (AAO) for men and women, respectively. We confirmed the normality assumption for age of onset variable for whole sample, and men and women separately using Shapiro-Wilk test. We used a conditional likelihood for the incident data as the covariates in this case are not biased and do not provide any extra information about the lifetime distribution, and the joint likelihood function for the prevalent data to incorporate the information from the biased distribution of covariates to the likelihood. The combined likelihood was a straightforward combination of both likelihood functions. To simplify interpretation, the age of onset (AAO) was centered at 80 years old. We created an interaction model by coding the two variables as follows:

$$Z_1 = \begin{cases} 1, & \text{for men} \\ 0, & \text{for women,} \end{cases} \tag{7.2}$$

$$Z_2 = \begin{cases} AAO - 80, & \text{for men} \\ 0, & \text{for women,} \end{cases} \tag{7.3}$$

and

$$Z_3 = \begin{cases} AAO - 80, & \text{for women} \\ 0, & \text{for men.} \end{cases}, \tag{7.4}$$

Under the coding, the baseline are women with onset of dementia at 80 years old. The regression parameters $\beta_1, \beta_2, \beta_3$ are the effect of gender, and years of onset centered at 80 years for men and women [7]. Since the biased distribution of covariates, $f_B(\mathbf{z})$ depends also on the regression coefficients $\beta_1, \beta_2, \beta_3$, we added these estimates to the table for comparison purpose. The results of maximum likelihood estimation of the relevant parameters are presented

in Table 7.2. Table 7.3 summarizes results from the estimation of regression parameters from the likelihood of covariates only.

Table 7.2: MLEs based on Weibull and log-normal $\mathcal{L}_{Prev}$, $\mathcal{L}_{Inc}$, and $\mathcal{L}_{Comb}$ for the CSHA data

| Weibull | | | | Log-normal | | | |
|---|---|---|---|---|---|---|---|
| *Parameter* | $\mathcal{L}_{Prev}$ | $\mathcal{L}_{Inc}$ | $\mathcal{L}_{Comb}$ | *Parameter* | $\mathcal{L}_{Prev}$ | $\mathcal{L}_{Inc}$ | $\mathcal{L}_{Comb}$ |
| $\hat{\alpha}$ | 1.49 | 2.68 | 1.74 | $\hat{\mu}$ | 1.61 | 1.58 | 1.66 |
| $\hat{\lambda}$ | 0.16 | 0.17 | 0.14 | $\hat{\sigma}$ | 0.62 | 0.57 | 0.60 |
| $\hat{t}_{0.25}$ | 2.71 | 3.69 | 3.49 | $\hat{t}_{0.25}$ | 3.29 | 3.30 | 3.51 |
| $\hat{t}_{0.50}$ | 4.89 | 5.13 | 5.79 | $\hat{t}_{0.50}$ | 5.00 | 4.85 | 5.26 |
| $\hat{t}_{0.75}$ | 7.78 | 6.64 | 8.62 | $\hat{t}_{0.75}$ | 7.60 | 7.13 | 7.88 |
| | | | | | | | |
| $\hat{\mu}_M$ | 1.52 | 2.58 | 1.77 | $\hat{\mu}_M$ | 1.41 | 2.58 | 1.67 |
| $\hat{\sigma}_M$ | 8.00 | 7.10 | 7.78 | $\hat{\sigma}_M$ | 8.00 | 7.10 | 7.80 |
| $\hat{\mu}_F$ | 4.00 | 5.60 | 4.51 | $\hat{\mu}_F$ | 4.10 | 5.60 | 4.50 |
| $\hat{\sigma}_F$ | 7.36 | 6.25 | 7.14 | $\hat{\sigma}_F$ | 7.36 | 6.25 | 7.14 |
| | | | | | | | |
| $\hat{\beta}_1$ | -0.24 | -0.17 | -0.23 | $\hat{\beta}_1$ | -0.25 | -0.08 | -0.22 |
| $\hat{\beta}_2$ | -0.04 | -0.02 | -0.04 | $\hat{\beta}_2$ | -0.04 | -0.02 | -0.03 |
| $\hat{\beta}_3$ | -0.03 | -0.02 | -0.04 | $\hat{\beta}_3$ | -0.04 | -0.02 | -0.04 |
| $\hat{p}$ | 0.33 | 0.3 | 0.32 | $\hat{p}$ | 0.33 | 0.3 | 0.32 |

Table 7.3: MLEs based on the maximum likelihood function of covariates only for the CSHA data

| *Parameter* | $\mathcal{L}_{Cov}$ |
|---|---|
| $\hat{\mu_M}$ | 2.58 |
| $\hat{\sigma_M}$ | 7.67 |
| $\hat{\mu_F}$ | 5.60 |
| $\hat{\sigma_F}$ | 7.00 |
| | |
| $\hat{\beta}_1$ | -0.35 |
| $\hat{\beta}_2$ | -0.60 |
| $\hat{\beta}_3$ | -0.07 |
| $\hat{p}$ | 0.32 |

Also, as mentioned before, the covariates in the incident sample do not contribute any information to the likelihood function and can be omitted from calculations. The parameters: $p, \mu_F, \sigma_F, \mu_M, \sigma_M$ in the incident sample can be obtained directly from the data.

Note that $\boldsymbol{\beta}\mathbf{z} = \beta_1 z_1 + \beta_2 z_2 + \beta_2 z_3$ and that $\mu(\boldsymbol{\theta})$ for the length-biased Weibull distribution can be calculated as follows:

$$\begin{aligned}
\mu_{Weib}(\boldsymbol{\theta}) &= \int_{\mathbf{z}} \mu(\mathbf{z};\boldsymbol{\theta}) f_z(\mathbf{z}) dz = \frac{\Gamma(1+\frac{1}{\alpha})}{\lambda} \int_{\mathbf{z}} e^{\boldsymbol{\beta}\mathbf{z}} f_{\mathbf{z}}(\mathbf{z}) dz \\
&= \frac{\Gamma(1+\frac{1}{\alpha})}{\lambda} \int_{gender} \int_{AAO} e^{\boldsymbol{\beta}\mathbf{z}} f_{z_1,z_2}(z_1,z_2) dz_1 dz_2 \\
&= \frac{\Gamma(1+\frac{1}{\alpha})}{\lambda} \int_{gender} \int_{AAO} e^{\boldsymbol{\beta}\mathbf{z}} f_{z_2|z_1} f(z_1) dz_2 dz_1 \\
&= \frac{\Gamma(1+\frac{1}{\alpha})}{\lambda} \int_{z_1} f(z_1) \int_{z_2} e^{\beta_1 z_1 + \beta_2 z_2} \Phi(\mu_M, \sigma_M) dz_2 dz_1 \\
&= \frac{\Gamma(1+\frac{1}{\alpha})}{\lambda} \int_{z_1} f(z_1) e^{\beta_1 z_1} \int_{z_2} e^{\beta_2 z_2} \Phi(\mu_M, \sigma_M) dz_2 dz_1 \\
&= \frac{\Gamma(1+\frac{1}{\alpha})}{\lambda} \int_{z_1} f(z_1) e^{\beta_1 z_1} e^{\beta_2 \mu + 0.5\sigma^2 \beta_2^2} dz_1 \quad \text{(by m.g.f of } z_2\text{)} \\
&= \frac{\Gamma(1+\frac{1}{\alpha})}{\lambda} p e^{\beta_1 z_1 + \beta_2 \mu_M + 0.5\sigma_M^2 \beta_2^2} + (1-p) e^{\beta_3 \mu_F + 0.5\sigma_F^2 \beta_3^2}.
\end{aligned} \tag{7.5}$$

Similarly, $\mu(\boldsymbol{\theta})$ for the length-biased log-normal distribution can be written as:

$$\begin{aligned}
\mu_{LogN}(\boldsymbol{\theta}) &= \int_z \mu(\mathbf{z};\boldsymbol{\theta}) f_z(\mathbf{z}) dz = \int_z e^{\mu + 0.5\sigma^2} e^{\boldsymbol{\beta}\mathbf{z}} f_z(\mathbf{z}) dz \\
&= e^{\mu + 0.5\sigma^2} \left[ p e^{\beta_1 sex + \beta_2 \mu_M + 0.5\sigma_M^2 \beta_2^2} + (1-p) e^{\beta_3 \mu_F + 0.5\sigma_F^2 \beta_3^2} \right].
\end{aligned} \tag{7.6}$$

Based on the parametric simulations (see previous section), we expected that the parameters obtained from all three likelihood functions will be very close. We also expected that the Weibull and log-normal survival curves from the combined population will be approximately between the incident and prevalent survival curves. The estimation results show that both parameters $\alpha$ and $\lambda$ for the Weibull distribution in the prevalent and combined cases are close.

While $\lambda_{inc}$ is close to $\lambda_{prev}$ and $\lambda_{comb}$, the shape parameter $\alpha_{inc} = 2.4$ in the incident sample is almost twice as big as the same parameter in the prevalent sample $\alpha_{prev} = 1.24$. Therefore, the individuals in the incident sample tend to have a higher probability of survival at smaller times, but their probability of survival decreases faster than for the individuals in the prevalent sample. However, we have to remember that the Kaplan-Meier estimator is not reliable for later survival times in a presence of heavy censoring (a small risk set for later times) as it tends to underestimate the variance of these times, and at the same time affect the shape parameter $\alpha$. We can see this easily if we look at the $\alpha$ parameters as a rough estimation of standard deviation of a sample expressed as $\alpha = 1/\sigma$. Small values of $\sigma$ cause the $\alpha$ parameter to blow up for the incident cases. The situation is a bit different for the prevalent cases; here $\alpha$ parameter is most likely underestimated due to the application of inverse weights to the survival times and therefore shifting the hump of the distribution toward 0. It is also interesting to note that for values of $\alpha$ close to 1, the probability of failure (the hazard function) becomes constant (as in the prevalent cases), while for values of $\alpha$ greater than 1, increases approximately proportionally to time (as in the incident cases). Since the people in the incident sample tend to be older than those in the prevalent sample, the failure rate increase much faster for the same survival times in the incident cases.

The estimation seems to improve in the case of log-normal distribution. Here, the $\mu$ and $\sigma$ parameters are close to each other as expected based on our simulation study. It may suggest that the parameters of the log-normal distribution are less sensitive to factors influencing shape of the survival curve than the Weibull's parameters. Other population characteristics such as gender, age at onset, type of dementia, or sampling place (institution or community) could possibly explain the differences we saw in the Weibull's shape parameter $\alpha$.

Since in the combined population we have approximately twice as much prevalent cases as incident cases, the combined likelihood function carries more weight from the prevalent part of the likelihood resulting in the estimates that are much closer to the estimates from the prevalent data than incident data. Therefore, reweighing the combined likelihood could be of future research interest. The fit of survival curves from the combined Weibull and combined log-

normal distributions to the nonparametric incident and prevalent survival functions represent an informal check of adequacy of both distributions and are shown in the Figure 7.2 and Figure 7.3.

Figure 7.2: Survival functions based on prevalent, incident, and combined Weibull likelihoods estimates for the CSHA sample

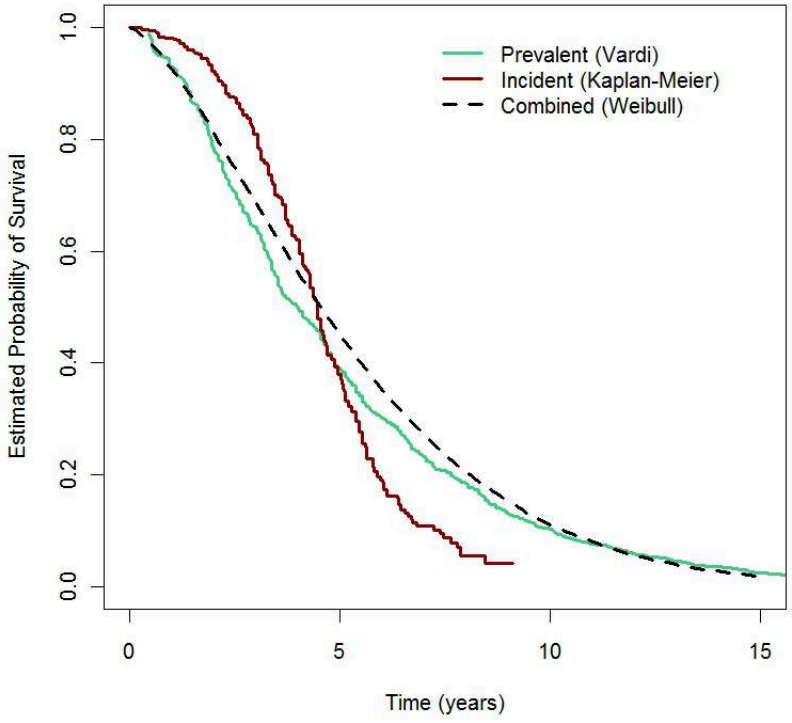
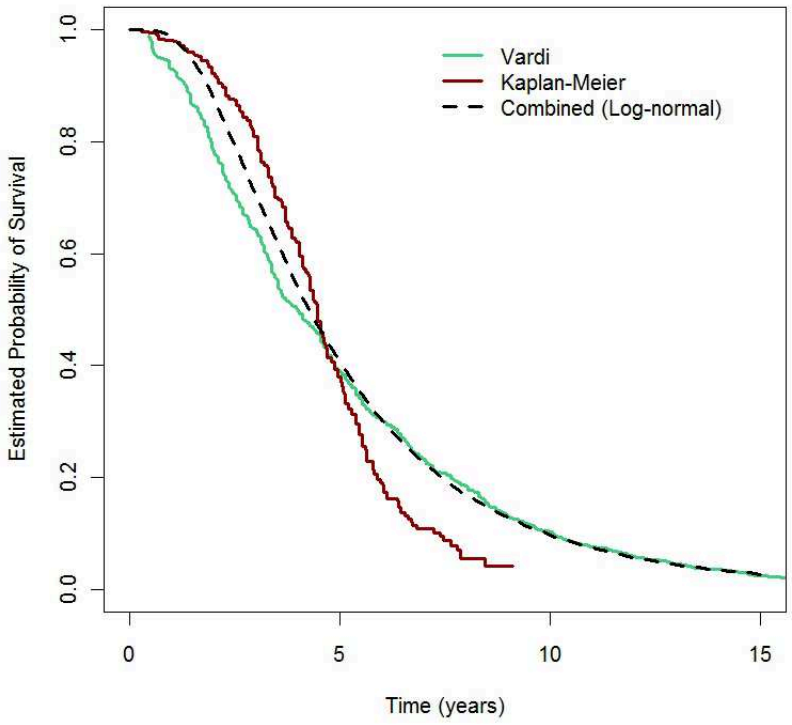


Figure 7.3: Survival functions based on prevalent, incident and combined Log-normal likelihoods estimates for the CSHA sample



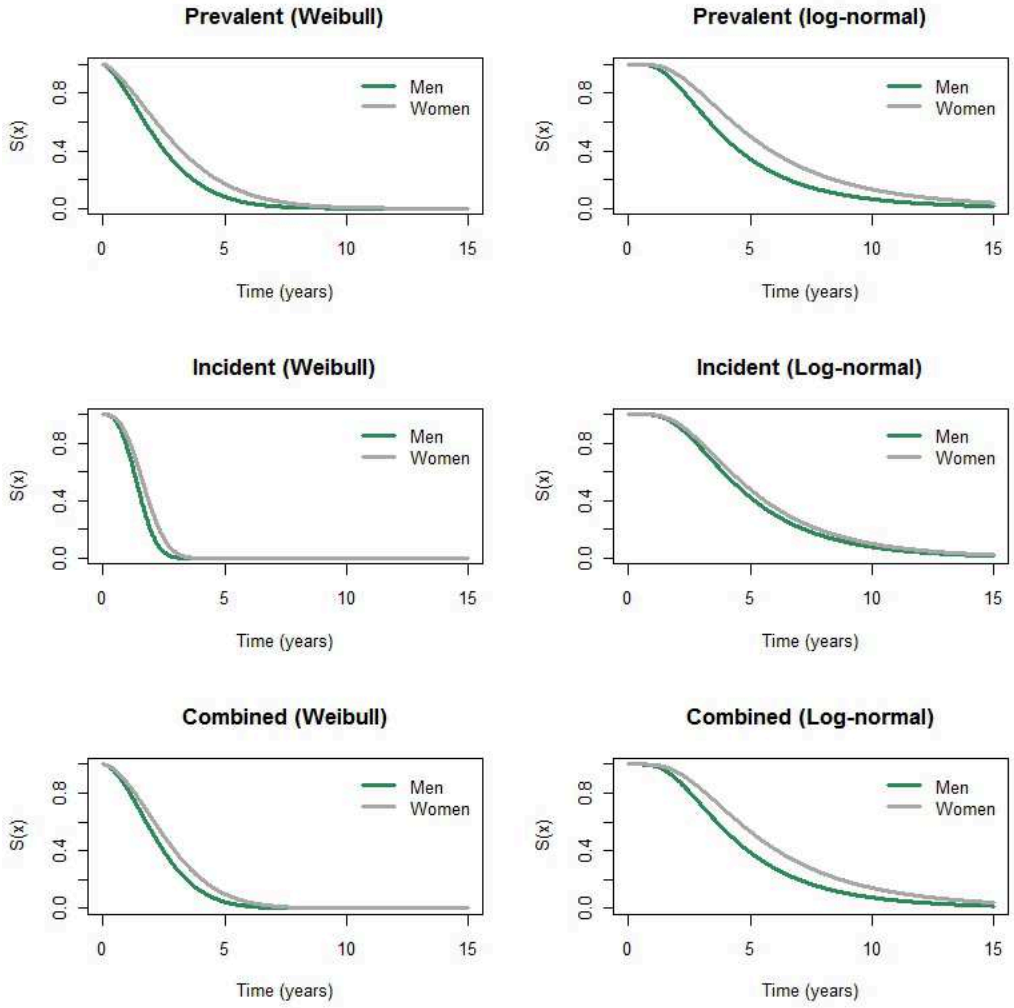

## 7.3 What doctors would like to know

In this section, we will look at the obtained estimates of the parameters and regression coefficients $\beta$s from a practical point of view.

The estimated time ratio (TR), a deceleration factor of survival time for any fixed value of $S(t)$ for the subpopulation with covariate $Z = 1$ compared to the baseline population ($Z = 0$), for the incident, prevalent, and combined populations for men vs. women was respectively 0.84, 0.78, and 0.79 in Weibull, and 0.92, 0.78, and 0.80 in log-normal distributions. It is interesting to note that the TR obtained from the covariates distribution only for the same baseline population was approximately the same: 0.70. Therefore, men tend to have shorter survival time than women (Figure 7.4). This can also be assessed based on the fact that $\beta_1 < 0$

in all the models. We can also see that since $\beta_2$ is approximately equal to $\beta_3$, the effect of age at onset in both populations is close and the survival time slowly decreases with a year increase in age (starting at age 80).

Figure 7.4: Survival curves for men vs. women at age at onset centered at 80 in the prevalent, incident, and combined populations based on Weibull and log-normal distributions




In both, Weibull and log-normal models, the average age of onset was approximately three years higher for women than for men with the estimates from the combined MLEs falling between the prevalent and incident estimates: 81.5, 82.6, and 81.7 for men and 84, 85.6, 84.5 for women in the prevalent, incident, and combined sample, respectively. The estimates for the

incident sample were in line with the median age of onset reported in the British incident study on dementia in which 13,004 individuals were followed for 14 years: 83 years for men and 84 years for women [47]. Note that although we used mean age of onset value, we assume that they are close to the median age of onsets since the normal model was used in their estimation [7]. Our incident estimate for men, and the combined estimate for women was the closest to the ones in the British study. Also, the results agree with the results from the study of the same CSHA data for prevalent cases only [7]. It is not a surprise in a case of the estimates from the prevalent data; however, the estimates of average age of onset from the combined likelihood were even closer to the one reported in the study: 82 years for men and 84 years for women.

For the Weibull model, the median survival times for men and women with the age of onset at 80 years were 84.41 and 85.25; and 83.88 and 84.98 years in the incident population and the prevalent population, respectively. In the combined population, the difference in the median survival years between women and men was slightly bigger than in the two samples separately: 85.70 and 84.5 years. For the log-normal model, the median survival times for men and women with the same age at onset were 84.5, 83.9, and 84.2; and 84.5, 85, and 85.3 for the incident, prevalent, and combined samples. The British incident study reported and overall median lifetime of 4.1 years for men and 4.6 for women. Again, our estimates were not far from these results, and in the case of the Weibull estimates for women, the combined likelihood agreed the most. The log-normal distribution performed sightly better in the estimation of the median survival time for both men and women (the incident and combined estimates closely resembled the one from the incident study).

Overall, both models, Weibull and log-normal perform similarly in estimation of the regression parameters in the prevalent, incident, and combined samples. The discrepancy between the estimates seem to be slightly smaller for the log-normal distribution. In the case, where the differences are not significant, the precision of the estimates should be taken in to the account. As we proved earlier, the increased efficiency of the combined MLEs makes it a preferable choice.

# Chapter 8

# Where to go from here?

Incident and prevalent studies are one of the most common epidemiological studies of natural history of a disease. While analyzing incident data provides the best quality information on history of a particular disease, for prevalent data, the bias introduced by sampling strategy, has to be account for to draw valid conclusions. In many longitudinal studies such as the Canadian Study on Health and Aging (CSHA) on dementia or the Medical Research Council's cognitive function and aging study (MRC CFAS), information is often collected from large number of individuals, both with and without disease, and over time both prevalent and incident cases are observed. Still, even for these studies, analyzes are often performed for either incident or prevalent sample. Utilizing all the data available and analyzing them simultaneously seem to be a natural method to improve inference and efficiency of estimates in these type of studies. In this thesis, we have explored a parametric combined likelihood function to obtain estimates based on the combined prevalent and incident datasets. In contrast to the likelihood function from prevalent and incident data alone, the combined likelihood utilizes all of the available information in a dataset, improves the power of statistical tests, and leads to less biased and more efficient estimates.

In order to illustrate and validate our approach, we used the CSHA prevalent and incident data as a benchmark for analysis and simulations. To be able to combine the unbiased (incident) data with a length-biased (prevalent) data, we first had to verify that both samples come

from the same population using a two-sample Smirnov type test. Otherwise, the combined approach would not be feasible and inferences drawn based on the MLEs from the combined likelihood function not reliable. In all our simulations, we used Weibull and log-normal distributions, two the most common distributions in survival analysis which proved to fit the CSHA data well [48]. We saw that, although not significant, the distances between the survival curves in both populations were larger than we would initially expect. We hypothesize that adjusting for a double bias originated from the exclusion of sampling weights and other important covariates from the analysis would decrease the observed distances between the survival curves. Further investigation of the properties of our test statistic in terms of power analysis for different censoring and sample size scenarios would be advisable.

There are also several issues with the NPMLE estimators we used in our work, which, if eliminated, could lead to more precise estimates of the survival, and in consequence, improve the performance of our "ad hoc" test. The KM estimator, although reliable in many scenarios, does not perform well in presence of large amount of censored data leading to estimates that are not reliable for larger survival times. On the other hand, the Vardi's NPMLE, not only tends to underestimate the survival for smaller times (due to inverse weighting method used to adjust for the length bias), but also is restricted to the prevalent data for which the assumption of stationarity holds. If a uniform distribution of truncation times cannot be assumed, the Vardi approach will not be valid. Therefore, a modification of the estimators, and extension of the Vardi's method to other sampling scenarios, as well as inclusion of covariates and sampling weights into the algorithm would be desired.

When one wants to improve estimation of the survival by incorporating covariates, two major regression models exist in the literature: the AFT and PH models. We chose to use the AFT models since they suit the parametric approach better, and their properties are preserved when the length bias is introduced to data. When dealing with length-biased data with covariates, one has a choice of conditioning on the values of covariates or incorporating the information included in their distribution by utilizing the joint likelihood function. Since it was shown that in the case of a length-biased data, covariates can provide an extra information about the survival

of subjects, we decided to use the joint approach in our analysis [5]. Because our approach to the combined incident and prevalent data was relatively new and the asymptotic properties of the MLEs are yet to be developed, we used simulations studies to validate our approach, compare efficiencies of the new likelihood with the ones from the prevalent and incident likelihood functions, and assess behavior of the combined likelihood in estimating the parameters of interest. Asymptotic efficiency and unbiasedness, as well as the behavior of the estimator for finite samples should be further investigated. Using parametric models poses itself a few limitations, and can be applied only when there is a strong indication that data follow a specific distribution. Therefore, other approaches, already existing in literature, such as nonparametric methods of estimation of survival function based on the combined likelihood should rather be considered.

Finally, in our study, we considered two important factors of the survival of dementia patients, gender and age at onset. We are aware of the fact that other covariates such as sampling place (institution vs community) or health factors could significantly impact the survival of subject and should be considered in the future analysis. The estimation of the age at onset of a disease in the prevalent data is an additional issue. Most of the studies estimate the date retrospectively from information gathered from family members and/or other informants. Other methods such as the Wolfson algorithm for calculating AAO from series of questions regarding subjects' past and present memory problems also exist in the literature [44]. Further improvement of the existing methods of estimating time of disease onset would lead to more accurate estimation of incident cases and ultimately allow obtaining more reliable estimates. Investigation of the effect of using different AAOs on the survival could be a subject of future research. In a long run, it would help to address some of the issues in the data collection process and design more effective studies.

Our work is a good first step to develop methods that would improve statistical estimation and inference from cross-sectional studies by combining information from both prevalent and incident data. Although our methods need improvement, we have shown that estimates of survival obtained from the combination of both type of data are more efficient and reliable

than those from incident or prevalent cohort studies alone. Filling the gap in the existing methods of analyzing and estimating the survival from longitudinal data should be of interest to both statisticians and practitioners. In addition, formalizing the approach by developing a ready to use comprehensive R package that would allow statisticians and non-statisticians use the proposed methods in relatively easy fashion would help not only popularize the methods, but also improve the overall methodology used in these type of studies. The importance of improving the methods proposed in this thesis is dictated by a growing number of studies such as the Canadian Longitudinal Study on Aging or the National Longitudinal Survey of Children and Youth that call for new, more efficient, and reliable estimation methods.

# Bibliography

[1] Aalen, O. O. (1972). “Nonparametric inference in connection with multiple decrement models.” Statistical Research Report no. 6, Department of Mathematics,University of Oslo.

[2] Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). “Statistical models based on counting processes.” Springer-Verlag, New York.

[3] Ashgarian, M., Wolfson, D.B. and Zhang, X. (2006). “Checking stationarity of the incidence rate using prevalent cohort survival data.” Stat Med.,25 (10), 1751-67.

[4] Asgharian, M. and Wolfson, D.B.(2005). “Asymptotic behaviour of the NPMLE of the survivor function when the data are length-biased and subject to right censoring. The Annals of Statistics, 33, 2109-2131.

[5] Asgharian, M., MLan, C.E. and Wolfson, D. (2002). “Length-biased sampling with right censoring: An unconditional approach.” Journal of the American Statistical Association, 97(457), 201-209.

[6] Bagdonavicius, V. and Nikulin, M. “Accelerated Life Models Modeling and Statistical Analysis.” (2001). Chapman and Hall.

[7] Bergeron, P.-J. and Cook, R.J.(2011). “Information in the sample covariate distribution in prevalent cohorts.” Statistics in Medicine, 30, 1397-1409.

[8] Bergeron, P.-J., Ashgarian, M. and Wolfson, D.B. (2008). "Covariates bias induced by length-biased sampling of failure times." Journal of the American Statistical Association, 103(482), 737-742.

[9] Bergeron, P.-J. (2006). "Covariates and length-biased sampling: Is there more than meets the eye?" PhD thesis, McGill University Department of Mathematics and Statistics.

[10] Bergeron, P.-J. (2003). "Measuring dependence using information gain when data are length-biased and right-censored." Unpublished MSc thesis, McGill University Department of Mathematics and Statistics.

[11] Bilker, W. and Wang, M.-C. (1991). "Bootstrapping left truncated and right censored data." Communications in Statistics-Simulation and Computation, 26(1), 141-171.

[12] Burton, A., Douglas, G.A., Royston, P. and Holder, R.L. (2006). "The disign of simulation studies in medical statistics." Statistics in Medicine, 25, 4279-4292.

[13] Caron, C., Asgharian, M. and Wang, M.-C. (2009). "Nonparametric incidence estimation from prevalent cohort survival data." Cobra Preprint Series, 54, 1-31.

[14] Chen, Y.Q., Nicholas, P. and Jewell N.P, and Yang, J. (2002). "Accelerated Hazards Model: Method, Theory and Applications." Available at: http:\\works.bepress.com \nicholas _ jewell \4

[15] Correa, J.A. and Wolfson, D.B. (1999). "Length-bias: some characterizations and applications." J.Statist. Comput. Simul., 64, 209-219.

[16] Cox, D. R. (1969). "Some sampling problems in technology." In New Developments in Survey Sampling. Edited by Johnson & Smith. Wiley.

[17] Cox, D.R. and Oakes D. (1984). "Analysis of survival data." Chapman & Hall/CRC, Boca Raton.

[18] Canadian Study of Health and Aging working group. (1994). "Canadian Study of Health and Aging: study methods and prevalence of dementia." CMAJ, 150, 899-913.

[19] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). "Maximum likelihood from incomplete data via the $EM$ alghorithm." Journal of the Royal Statistical Society, 39(1), 1-38.

[20] Efron, B. (1981). "Censored data and the bootstrap." Journal of the American Statistical Association, 76(374), 312-319.

[21] Ellenberg, S.S. and Siegel, J.P. (1997). "Survival analysis in the regulatory setting." In Proceedings of th First Seattle Symposium in Biostatistics: Survival Analysis, Eds. D.Y. Lin and T.R. Fleming, pp.xxx-xxx. Springer-Verlag, New york.

[22] Fleming, T.R. and Harrington, D.P. (1991). "Counting processes and survival analysis." John Wiley & Sons, New York.

[23] Good, P.I and Hardin, J.W. (2009). "Common errors in statistics (and how to avoid them)." John Wiley & Sons, New Jersey.

[24] Graunt, J.(1662). Natural and political observations made upon the bills of mortality, 1662.

[25] Hjort, N. L. (1992). "On inference in parametric survival data models." International Statistical Review, 60, 355387.

[26] Huang, Ch.-Y. and Qin, J. (2011). "Nonparametric estimation for length-biased and right-censored data." Biometrika, 98(1), 177-186.

[27] Huang, M.L. (2008). "A weighted estimation method for survival function." Applied Mathematical Sciences, 2(16), 753-762.

[28] Kaplan, E. L. and Meier, P. (1958). "Nonparametric estimation from incomplete observations." Journal of the American Statistical Association, 53, 457-481.

[29] Krickerberg, K. (2005). "Survival analysis : A self-learning text." Springer, New York.

[30] Lai, T.Z. and Ying, Z. (1991). "Estimating a distribution function with truncated and censored data." The Annals of Statistics, 19(1), 417-442.

[31] Lawless, J.F. (2003). "Statistical models and methods for lifetime data." John Wiley & Sons, New Jersey.

[32] Lee, T. and Wang, J.W. (2003). "Statistical method for survival data analysis." John Willey & Sons Inc., New Jersey.

[33] Mantel, N. (1966). "Evaluation of survival data and two new rank order statistics arising in its consideration." Cancer Chemotherapy Reports, 50 (3), 16370.

[34] McCullagh, P. (2005). "Proportional-Odds Model. Encyclopedia of Biostatistics."

[35] Mitchell, B. (1971). "A comparison of chi-square and Kolmogorov-Smirnov tests." Area, 3(4), 237-241.

[36] Moeschberger, M. L. and Klein, J. P. (1997). "Survival analysis." Springer, Secaucus, New Jersey.

[37] Nelson, W. (1969). "Hazard plotting for incomplete failure data." Journal of Quality Technology, 1, 27-52.

[38] Oakes, D. (2001). "Biometrika centenary: Survival analysis." Biometrika, 88(1), 99-142.

[39] Panchenko, D. (2006). "Statistics for applications." (Massachusetts Institute of Technology: MIT OpenCourseWare), http://ocw.mit.edu (Accessed 03 Jan, 2012). License: Creative Commons BY-NC-SA.

[40] Rao, C.R.(1977). "A natural example of a weighted distribution." Amer. Stat., 31, 24-26.

[41] Ross, S.-M. (2006). "Simulations". Elsevier Academic Press, London.

[42] Royston, P. (2001). "Flexible parametric alternatives to the Cox model, and more." The Stata Journal, 1(1), 1-28.

[43] Tsai, W.-Y., Jewell, N. P. and Wang, M.-C. (1987). "A note on the product-limit estimator under right censoring and left truncation." Biometrika, 74, 883-886.

[44] Wolfson, C. and Rouah, F. (2001). "A Recommended Method for Obtaining the Age at Onset of Dementia From the CSHA Database." International Psychogeriatrics, 13(S1), 57-70.

[45] Vardi, Y. (1982). "Nonparametric Estimation in the Presence of Length Bias." Ann. Statist., 10 (2), 616-620.

[46] Vardi, Y. (1989). "Multiplicative censoring, renewal processes, deconvolution and decreased density: Nonparametric estimation." Biometrika, 76(4), 751-61.

[47] Xie, J., Carol, B. and Matthews, F.E. "Survival times in people with dementia: analysis from population based cohort study with 14 year follow-up." (2008). BMJ: British Medical Journal (International Edition), 336(7638), 258-62.

[48] Younger, J. (2012). "Goodness-of-fit for length-biased survival data with right-censoring." MSc thesis. University of Ottawa Department of Mathematics and Statisitcs.

[49] Zelen M. and Feinleib M. (1969). "On the theory of screening for chronic diseases." Biometrika, 56, 601-614.