Covariates and length-biased sampling: Is there more than meets the eye?

Pierre-Jérôme Bergeron

Department of Mathematics and Statistics

McGill University, Montréal

September 2006

A thesis submitted to the Faculty of Graduate Studies and Research

in partial fulfilment of the requirements of the degree of

Doctor of Philosophy

Your file Votre référence
ISBN: 978-0-494-32148-5
Our file Notre référence
ISBN: 978-0-494-32148-5

Abstract

It is well known that when subjects with a disease are identified through a cross-sectional survey and then followed forward in time until either failure or censoring, their estimated survival function of the true survival function from onset are biased. This bias, which is caused by the sampling of prevalent rather than incident cases, is termed length bias if the onset time of the disease forms a stationary Poisson process. While authors have proposed different approaches to the analysis of length-biased survival data, there remain a number of issues that have not been fully addressed. The most important of these is perhaps that of how to include covariates into length-biased lifetime data analysis of the natural history of diseases, that are initiated by cross-sectional sampling of a population. One aspect of that problem, which appears to have been neglected in the literature, concerns the effect of length-bias on the sampling distribution of the covariates. If the covariates have an effect on the survival time, then their marginal distribution in a length-biased sample is also subject to a bias and is informative about the parameters of interest. As is conventional in most regression analyses one conditions on the observed covariate values. By conditioning on the observed covariates in the situation described above, however, one effectively ignores the information contained in the distribution of the covariates in the sample. We present the appropriate likelihood approach that takes into account this information and we establish the consistency and asymptotic normality of the resulting estimators. It is shown that by ignoring the information contained in the sampling distribution of the covariates, one can still obtain, asymptotically, the same point estimates as with the joint likelihood. However, these conditional estimates are less efficient. Our results are illustrated using data on survival with dementia; collected as part of the Canadian Study of Health and Aging.

Résumé

Il est bien connu que lorsque les sujets malades sont identifiés à partir d'une étude à coupe transversale et ensuite suivis jusqu'à défaillance ou censure, la fonction de survie estimée est biaisée par rapport à la vraie fonction de survie à partir du temps de début. Ce biais, dû à l'échantillonnage de cas prévalents au lieu de cas incidents, est nommé biais de longueur lorsque les temps de début de la maladie forment un processus de Poisson stationnaire. Bien que des auteurs aient proposé plusieurs approches pour analyser des données à biais de longueur, il reste plusieurs problèmes qui n'ont pas été complètement étudiés. Le plus important d'entre eux est possiblement comment inclure des covariables dans l'étude de durées de vie dans l'histoire naturelle de maladies, qui est initiée par l'échantillonnage à coupe transversale d'une population. Un aspect de ce problème, qui semble avoir été négligé dans la littérature, concerne l'effet du biais de longueur sur la distribution des covariables dans l'échantillon. Si les covariables ont un effet sur la durée de vie, leur distribution marginale dans un échantillon à biais de longueur est également affectée par un biais qui comporte de l'information sur les paramètres étudiés. Habituellement, dans la plupart des analyses de régression on conditionne sur les valeurs des covariables. Cette approche conditionnelle dans la situation décrite ci-dessus fait fi de l'information contenue dans la loi marginale des covariables de l'échantillon. On présente une approche par vraisemblance jointe qui tient compte de cette information et on établit la consistence ainsi que la normalité assymptotique des estimateurs qui en découlent. On démontre que prendre une approche qui ne tient pas compte de l'information contenue dans la distribution des covariables donne assymptotiquement les mêmes valeurs ponctuelles estimées. Cependant, ces estimateurs conditionnels sont moins efficaces. Les résultats sont illustrés avec des données de durée de vie avec démence, recueillies à partir de l'Étude canadienne sur la santé et le vieillissement.

Acknowledgments

There are many people I wish to thank. Without their help and support, this thesis would not have been possible. I would like to thank above all my supervisor Professor Masoud Asgharian. His advice and contagious enthusiasm have helped me greatly throughout this endeavour. I would also like to thank Dr. Russell Steele for his insight with regards to the computational issues of this thesis.

On a less academic and more personal level, I would like to give special thanks to my family for their support and encouragement, and to my girlfriend Eugenie for her love, support and kindness. She must be credited for drawing the first two figures in this thesis.

This work was supported in part by grants from Fonds québécois de la recherche sur la nature et les technologies and the National Sciences and Engineering Research Council of Canada. The data reported in this article 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) and no. 34 (96-1)). The study was coordinated through the University of Ottawa and the Division of Aging and Seniors, Health Canada.

Contents

Chapter 1

Introduction

The real life statistical problems one can be faced with often differ in many ways from the ideal setting of the basic, college level statistics course. A number of the standard assumptions from the sanitized realm of textbook problems fail to hold in modern applications. For example, the analysis of lifetime data has its own set of common peculiarities. One of them is that the variable of interest, time to an event, is a positive random variable that is rarely even approximately normally distributed. Another common feature in this context is censoring: the variable of interest may not be fully observed. It is often impractical, due to time and logistical constraints, to follow every single subject to observe the event of interest. That is why information collected on some subjects is inevitably incomplete. By now, there is a vast literature and a number of tools to address statistical inference based on censored data. The ideal setting for failure time data is the so-called incident studies. The following diagram depicts an incident study.

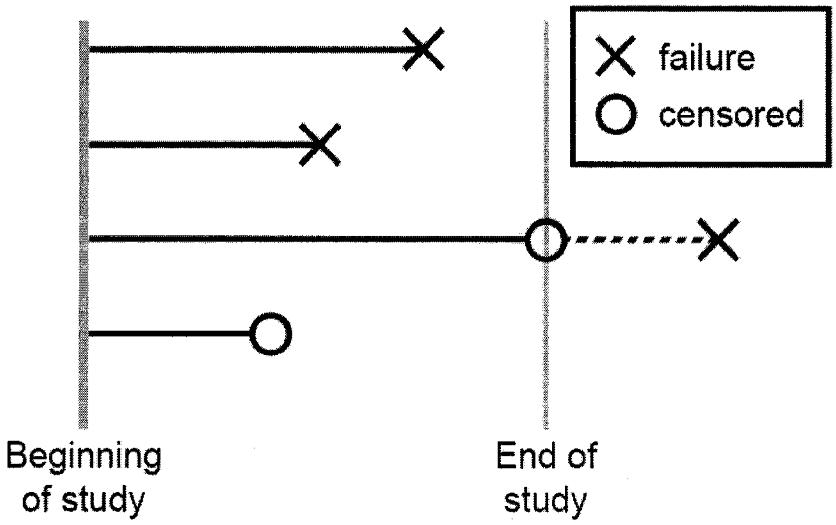

Figure 1.1: Incident study

In such studies, subjects are recruited or observed from the time of initiation of an event, such as onset of a disease, or turn of a switch for light bulbs, and followed until censoring or occurrence of the event. Censoring can occur at the end of a study but is not limited to this scenario. Human subjects, for example, can be lost to follow-up. This is illustrated in Figure 1.1, the bottom subject is lost at some time in the study, while the one above would have been observed had the study been longer. In such a setting, the tools available (such as the Kaplan-Meier estimator) rely on the assumption that censoring is noninformative about the survival time.

In the case of rare diseases or when other constraints prevent the recruitment of incident cases, the alternative approach is the study of a prevalent cohort, collected often through cross-sectional surveys. Under this scenario, the subjects have already experienced the initiation of the event prior to entering the study (this is called left truncation) and they are followed until failure or censoring.

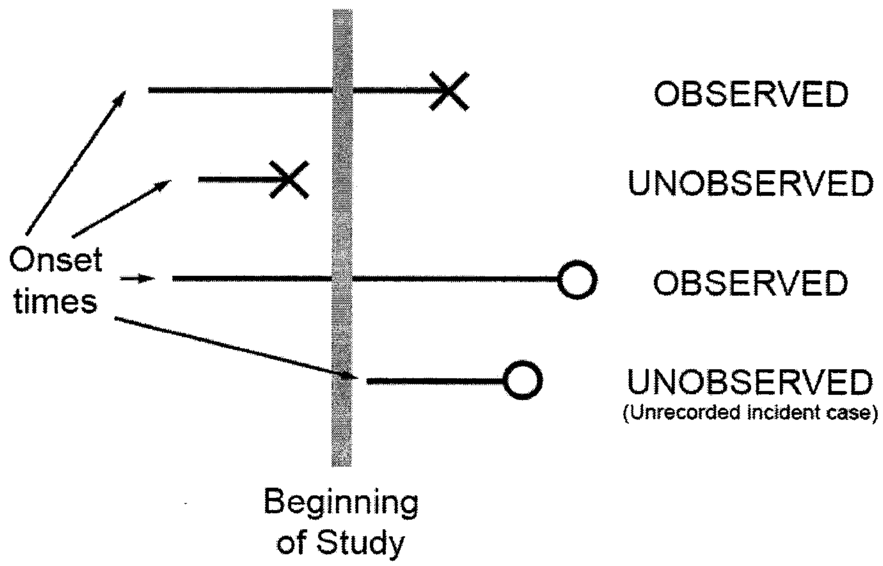

Figure 1.2: Prevalent study

This is no longer the ideal setting, as longer lived subjects are more likely to be selected into the study while the shorter lived one may go unobserved. This means that prevalent cases are not representative of the incident population. When interest lies in estimating the survival distribution, from onset, of subjects with the disease one must take into account that the lifetimes ascertained in such a fashion are left-truncated. Additionally, as the time from onset to recruitment is contained in both the full lifetime and censoring time, the censoring times are informative and the methods applicable to incident cases cannot be used on prevalent cases without adjustment.

Sampling from a prevalent population is a form of biased sampling. When it is possible to assume that there has not been an epidemic of the disease over the period of time that covers the onset times of the subject, one may assume that the incidence process of the disease is a segment of a stationary Poisson process. Under this assumption, the sampled lifetimes are length-biased, that is, the probability of recruiting a subject is directly proportional to its lifetime. Figure 1.3 illustrates how the length-biased density shifts the weight of its associated unbiased density toward the higher values of the variable.

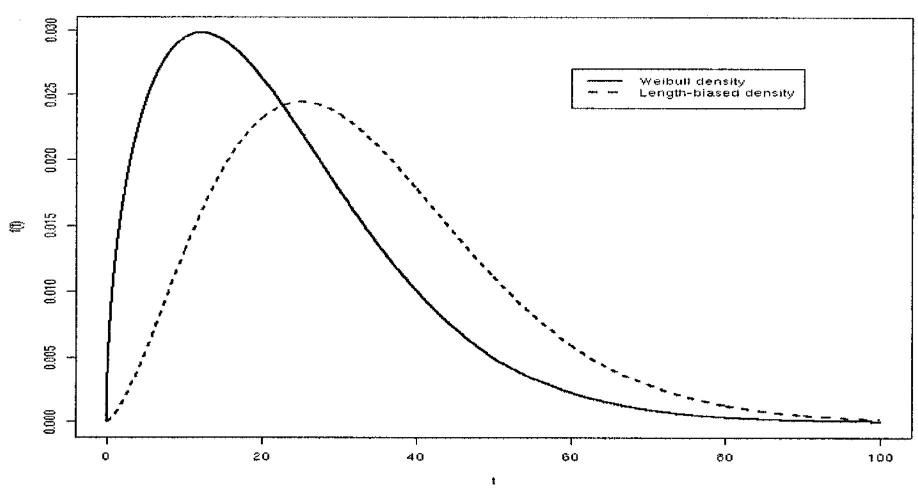

Figure 1.3: Unbiased Weibull density and associated length-biased density

Ignoring length bias leads to an overestimation of the survival function, as shown in Figure 1.4

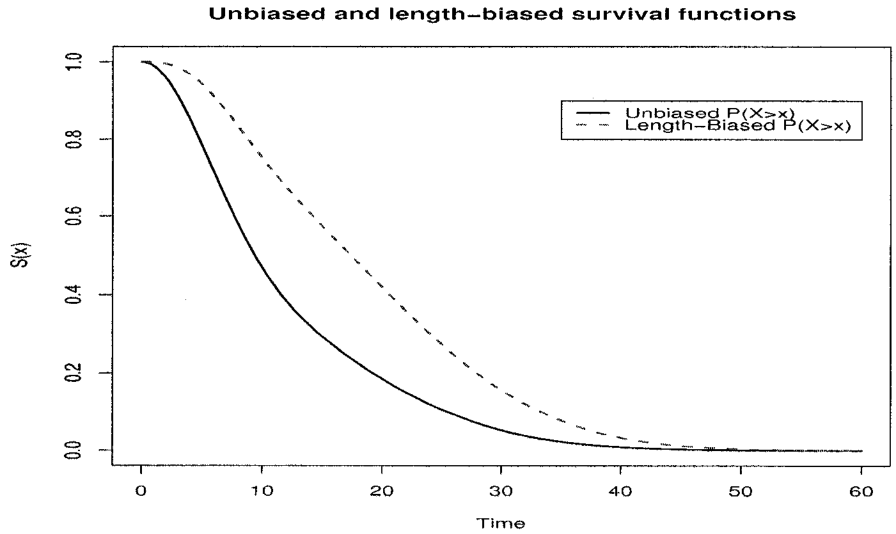

Figure 1.4: Unbiased and length-biased survival functions

The following example illustrates the concept of length-biased sampling. Suppose a graduate student walks by the Schulich School of Music everyday at a random time distributed uniformly in a 9 minute interval, the time at which a musician practices Bach's Toccata and Fugue in D minor, BWV 565. So, every day, the student will hear part of this well known musical composition. In most interpretations the Toccata lasts about three minutes while the Fugue is around six minutes long, so the graduate student will be twice as likely to hear part of the Fugue than part of the Toccata because the former occupies more space in the interval. Hence this is a case of length-biased sampling. Additionally, over time, the number of fugues heard by the student will be about double the number of toccatas, even though the musician plays one of each everyday. But how does this relate to the main point of this thesis?

While measuring the duration of musical pieces holds little scientific interest, the purpose of lifetime data analysis is often not only to describe survival distribution but to measure the impact of covariates on survival. When the sampling mechanism is subject to a bias relating to the time variable, it also induces a bias in the distribution of the covariates. Figure 1.5 shows how a mixture of two Weibull densities is affected by length-biased sampling: the proportion occupied by the subpopulation with higher mean is increased in the biased density. While this might appear natural as it is, the implications in regression analysis are more subtle. In standard regression, the variable of interest is modeled conditionally on the covariates, and the marginal distribution of the covariates is left out of the analysis because it holds no information on their impact. When the sampling is biased, this is no longer the case.

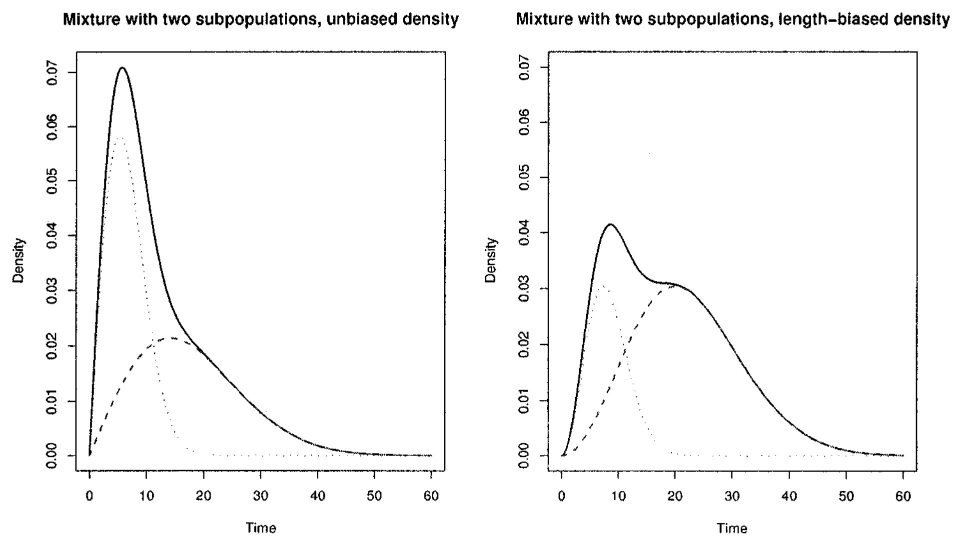

Figure 1.5: Unbiased and length-biased mixture of Weibull densities

To fully appreciate the impact of this, consider the following illustrative example. Suppose a graduate student decides to "reinvent the wheel" and makes a project of measuring the impact of gender on survival. Knowing a little bit of lifetime data analysis, the student knows the ideal setting is an incident study, which translates in the project as sampling individuals at birth. This poses some difficulties as obtaining permission to follow newborns from the maternity ward until their eventual death raises some privacy and ethical issues which would make the recruitment of subjects overly complicated. Additionally, the subjects would likely outlive the zealously motivated researcher, so if the project was ended after a few years, most if not all observations would be censored, yielding inconclusive results. Reluctantly, the student realizes that prevalent cases would make the sampling process much easier and reduce the project duration greatly. This would mean observations biased toward those who live longer, but this can be taken care of in the analysis. To find subjects who should die in a reasonable amount of time for the project, the researcher goes to a residence for the elderly and decides to visit every residant there, record their gender and follow them until their demise. The student quickly loses his motivation as he realizes that too many residents are really delighted to see a visitor, and that it would just be too depressing to wait for these new friends to die. So he goes home with only data on the gender of elderly subjects. Performing some basic tests on this rudimentary dataset, the researcher realizes that there are significantly more women than men at the "old folks home". This discrepancy cannot be the result of chance, as genetics have kept the distribution of gender at birth uniform. Also, the Charter of Rights and Freedoms prevents the residence from using discrimination based on gender to accept tenants. But one has to reach a certain age before being admitted into such a care center, which means that men having shorter lifetime could explain this imbalance. Hence, the student finds information about the effect of gender on survival simply by looking at the distribution of gender in a prevalent sample, *without collecting any lifetime data.*

Using a conditional approach in lifetime regression in the presence of length bias discards the information contained in the marginal distribution of the covariates. In this thesis, we propose a new joint likelihood approach for the analysis of length-biased, right-censored data that incorporates this information from the sampling distribution of the covariates. This thesis is organized as follows: in Chapter 2, we review the history of length-biased sampling and the analysis of survival data from a prevalent cohort. Because it is relevant to the subject, a section on different forms of biased sampling is included. In Chapter 3, we derive the likelihood function in the presence of covariates and expose the relationship between length-bias sampling for event-time and mean-lifetime bias for the covariates. We further compare the correct likelihood with the conditional likelihood that disregards the sampling distribution of the covariates, showing that both asymptotically result in the same point estimates. An entire section is devoted to an analytical example that shows that estimates obtained through the correct joint likelihood are more efficient than parameter estimates computed using the conditional likelihood. Chapter 4 focuses on the asymptotic properties of the maximum likelihood estimator. We establish the consistency of the MLE, and derive the asymptotic normality of the MLE. In Chapter 5, we discuss how to adapt two standard regression models in survival analysis to length-biased sampling. For the purpose of implementation, we propose new algorithms for estimation and simulation for the problem in Chapter 6. Chapter 7 is devoted to applications. The methods are tested on simulation studies and then applied to a set of real data to estimate the survival function of subjects diagnosed with dementia using data collected as part of the Canadian Study for Health and Aging (CSHA). The rest of the thesis concentrates on other research relating to the main subject; Chapter 8 presents a summary of less successful efforts in the domain, while Chapter 9 offers a recapitulation and a look at new avenues of research.

Chapter 2

An abridged history of length-biased lifetime data

In this chapter, we present an overview of the literature pertaining to the main problem addressed in this thesis, namely how to accommodate covariates in the analysis of length-biased data in the presence of right censoring as can occur in survival data analysis. We begin by discussing the first papers written on length-biased sampling, then go into how length-bias arises in survival data through cross-sectional sampling from prevalent cohort. We follow this with a discussion on the assumption of stationarity, how to assess its validity and what tools are available when this assumption fails. The next step is exploring the unconditional approach developed initially by Vardi which applies when stationarity holds and the data are properly length-biased. Finally, because incorporating covariates brings another layer of bias in the data, we review some of literature on more general biased sampling issues.

2.1 The early writings

It is interesting to note that a literature review, being a nonrandom process, is subject to biased sampling from the population of articles on the subject at hand: long-lived, well-known and popular journals are more likely to be browsed, and heavily cited articles will have a higher probability of being noticed. There is also time frame issue, at some point, older works either become common knowledge in the field or they simply disappear into obscurity. A new problem rarely comes with a name attached to it. Whoever stumbles upon it first might not use the same nomenclature that will develop and settle as the problem gains in popularity and a single problem might have multiple names depending on its field of application.

Hence there are two questions which can be difficult to answer. One is whether a problem has already been solved, or at least studied, which is a very important question for a doctoral candidate. The other question, given the problem has been identified, is who tackled it first? Even on the subject of length-biased sampling, one cannot discount the possibility of that, say, Archimedes might have worked on it. There is a remote chance that, concealed in a palimpsest buried in some medieval Byzantine monastery, solutions based on infinitesimals and geometry exist to answer why it appears that big ships burn more easily than small ones when targeted by the famous death ray that supposedly defended Syracuse from naval assault.

To the best of our knowledge, Wicksell (1925) was the first to mathematically describe and resolve a problem involving length-biased sampling. Particularly interesting is the setting of his research: anatomy. Wicksell was faced with what he called a "corpuscle problem": spherical objects (e.g. follicles) that vary in size are uniformly distributed within the volume of an opaque body (in this case, an organ, such as the spleen). The goal is to estimate the number and size distribution of the corpuscles contained within the body. With today's computers and MRIs, this would probably reduce to measuring densities inside the body and finding thresholds to isolate the follicles, but back in Wicksell's day such tools did not exist, so sampling had to be done through simpler means. Take the body, cut it in half, and look at the corpuscles that were sliced in the cross-section. There was also some magnifying and paper tracing involved but those details do not play a role in the analysis. This is length-biased sampling as spheres with bigger diameter are more likely to appear in the cut, but additionally the projected, apparent diameter is smaller than the true diameter. So the problem is three-dimensional but the data are in two dimensions, and the bias is coupled with some sort of truncation (different than what will be discussed later). Wicksell (1926) revisited this problem by extending his previous works to ellipsoids corpuscles.

The systematic theoretical study of length-bias began with McFadden

(1962). The concern was with the length of intervals in a stationary point process. Consider a random sequence of events $\ldots T_{i-1} \leq T_i \leq T_{i+1} \ldots$ where the distribution of events is invariant with respect to calendar time. One can define a random variable for the length of an interval between two consecutive events, sampled at a random event T_i to be

$$X_i = T_{i+1} - T_i. \tag{2.1.1}$$

Here all X_i have a common distribution function $F_1(x)$ because of stationarity. The study of this kind of X_i is just basic renewal process theory. Another interval length of interest is when one samples at a random time t. Now instead of just one interval, there are really two to look at: the time from the previous event to t, which we will call the *backward recurrence time* and denote $L_0 = t - T_0$; and the *forward recurrence time*, from t until the next even T_1, which will be denoted $L_1 = T_1 - t$. The interval length between the two event is given by

$$X_i^* = T_{i+1} - T_i = L_1 + L_0. \tag{2.1.2}$$

McFadden then proceeded to study the fundamental differences between X_i and X_i^* and their relationship. Note that here length bias is induced by sampling at a point in time and assessing the full length of the one-dimensional observation, setting quite different from Wicksell's but much more common in the literature.

Blumenthal (1967) studied the implications of "proportional sampling" in life length studies. This author wanted to estimate the mean life of electron tubes that are sampled while operating since some prior date, which means that those sampled tubes tend to live longer than the population intended for study. This is again a renewal process setting, as the model involves replacing a component that has failed by another one which has already been in use. The backward and forward recurrence times are noted $U(T)$ and $V(T)$, the sum of which is denoted $L(T)$. As the interest lies in estimation, since $U(T)$ and $V(T)$ are theoretically identically distributed but $U(T)$ is available immediately while measuring the remaining life time is inherently time consuming, Blumenthal ansked, "Why should one bother with obtaining $V(T)$?" His answer is that since $U(T)$ and $V(T)$ are not independent, using both $U(T)$ and $V(T)$ should give estimates with less variance. Blumenthal applied the theory to Gamma and Weibull life times, deriving some MLEs and moments for those distributions.

The last paper we consider in this section was by Cox (1969). He described what induces length-biased sampling (again, under the name of "proportional sampling") in industry. Like Wicksell's problem, those are generally technological issues. The particular experiment under scrutiny was the sampling of fibers in fabric, and Cox described the moments of the sampled lengths in terms of the unbiased (unweighted) distribution moments. Cox discussed estimation of parameters from unbiased and length-biased sampling with Log-Normal and Gamma distributions, showing that length-biased sampling gives more efficient estimates of the upper tail of the distributions. That is expected as this is where there will be more observations through the biased sampling mechanism. None of the aforementioned author considered censoring, as few dealt with applications where censoring can occur.

2.2 Length biased sampling: some applications

With the basic theory established, the applications multiplied over time. Goldsmith (1966) found that Wicksell's anatomy problem had equivalents in physics. He calculated true particle sizes observed from thin slices, working with discrete and continuous true size distributions. Similar issues arose in astrophysics on the problem of cataloging galaxies and biology on the subject of carcinogenesis (Neyman 1969). But length bias often occurs when the variable of interest or sampling mechanism involves time. In economics, while time might not be the main focus of a study, it often can play a role in sampling. Nowell, Evans and McDonald (1988) studied how length-biased sampling affects contingent value studies which are used to quantify the value of non monetary variables, such as environmental commodities and non traded goods. Sampling users of such commodities (e.g. a fishing resort) often involves being on site while individuals are in the middle of an activity, which means that those who spend more time and thus put more value on the contingents are more likely to be sampled. Nowell and Stanley (1991) observed similar length bias in mall intercept surveys. They also pointed out that while sampling shoppers inside a shopping center rather than at the entrance exhibits some properties of length bias, other bias mechanisms might be involved (e.g. individuals who visit many different stores will be sampled more easily than those who simply spend a lot of time in a single shop then leave).

When it comes to literature on medical and epidemiological data, length-biased sampling comes in more than one form. These applications can at least be traced back to Fisher (1934) and Neyman (1955) on prevalence-incidence bias. Zelen's work on screening tests in the late 60's and early 70's were landmarks in recognition of implications of biased sampling in medical and epidemiological studies. Davidov and Zelen (2001) studied relative risk and family history, showing how analyzing family registries will pick up observations proportionally to the family size. But the main concern of this thesis is survival data where time is the variable of interest. In this setting, length bias arises through the study of prevalent cohorts. Such data are left-truncated as subjects are sampled after the initiation of the event of interest. Applications of prevalent cohorts includes AIDS, studied by Lagakos, Barraj and De Gruttola (1988) (the data in this case are originally right-truncated but the authors use the reverse time to analyze it as left-truncated data), and Alzheimer's disease. Gao and Hui (2000) used two-phased sampling to estimate incidence of dementia. Stern et al. (1997) analyzed prevalent cohort data for Alzheimer but did not correct for length-bias. Wolfson et al. (2001) used data from the Canadian Study of Health and Aging (CSHA) and demonstrated that, when length-bias is taken into account, the median survival lifetime for individuals diagnosed with dementia is considerably shorter than was previously estimated.

2.3 The conditional approach and stationarity

While the early works on length-biased sampling emphasized parametric models, their theoretical developments have since mainly focused on nonparametric methods. One necessary assumption required for left-truncated data to be properly length-biased is stationarity of the process generating the onset times. When this assumption does not hold, there will still be a bias in the prevalent population compared to the underlying incident cases, but the tools based on purely length-biased data will not be valid. Brookmeyer and Gail (1987) described the relationship between the distribution of onset times and the prevalent-incident bias. Without the stationarity assumption, one can deal with the general problem of left truncation from a conditional perspective. This approach was developed initially by Turnbull (1976) as an extension of nonparametric product limit estimator proposed by Kaplan and Meier (1958). It was further investigated by Wang, Jewell and Tsai (1986) who derived the asymptotics of the product limit estimate in the presence of random truncation. Wang (1991) extended this to nonparametric maximum likelihood estimation of the truncation distribution. Other works in that area include Anderson, Borgan, Gill and Keiding (1993) and Tsai, Jewell and Wang (1987). Hazard regression for length-biased data has also been studied by Wang (1996), while Huang and Wang (1995) examined competing risk models for prevalent data. Some authors have forayed into semiparametric territory. Wang (1989) has investigated a semiparametric approach for randomly truncated data. Other related work has been done by Tsokidov (1998), Van der Laan and Hubbard (1998), and Gilbert, Lele and Vardi (1999). These works are all based on the so-called quasi-independence assumption. Martin and Betensky (2005) have proposed methods for testing quasi-independence between failure and truncation times using Kendall's tau. Zelen (2004) modeled recurrence time distributions for prevalent cases, and includes some generalization of length-biased sampling. Huang and Wang (2004) jointly modeled recurrence event process and failure times. Though the conditional approach enables one to work with left-truncation of unspecified distribution, it does not offer the most efficient estimates for the survival function when the stationarity assumption holds. Consequently this has led to some exploration of the subject. Asgharian, Wolfson and Zhang (2006) offered a characterization and informal test for stationarity, while Addona and Wolfson (2006) proposed the first formal test for the stationarity of incidence rate in prevalent cohort studies.

2.4 The unconditional approach

When the assumption of stationarity is reasonable or formally verified, the analysis of left truncated data can proceed through an unconditional maximum likelihood approach that allows one to recover the underlying incident and unbiased distribution of the lifetime. This approach was pioneered by Vardi (1982, 1985) and Gill, Vardi and Wellner (1988). Vardi (1989) derived the NPMLE (obtained through an EM algorithm) for length-biased right-censored data and Vardi and Zhang (1992) presented the asymptotics for this NPMLE. The setting of these papers however is not a realistic setting for prevalent cohort study with follow-up. Vardi's algorithm relies on fixed number of uncensored and censored observations (that is, whether a subject will be completely observed or not is known *a priori*, which is quite a strong assumption), and the author mentions that under other conditions such asymptotics will not hold and have to be derived separately, though the likelihood remains the same. This work was accomplished by Asgharian, M'Lan and Wolfson (2002) and Asgharian and Wolfson (2005), who have derived asymptotic results for the NPMLE of the length-biased and unbiased survival functions when the data are length-biased and subject to right censoring.

2.5 Covariates and other kinds of bias

When covariates are to be modeled, however, the ostensible choice is between semiparametric and fully parametric models such as the Weibull model. There are different settings in which length-bias must be taken into account and at the same time covariates allowed for. These include treatment studies (Wang et al. 1993) and natural history of disease studies (Alioum and Commenges, 1996; and Cnaan and Ryan 1989). In addition, Brookmeyer and Gail (1981) examined the effect of using prevalent cohorts on relative risk. However, there has been no satisfactory way proposed to use the Cox proportional hazards model when the natural history of a disease is of interest and the data are length-biased. The assumptions that must be made are rather restrictive. Wang (1996) showed how the Cox model may be fit to length-biased data and partial (quasi) likelihood inference made about the regression parameters when there is no censoring. For full discussion of these and related issues, see Addona (2001).

In the context of common regression analysis, when sampling does not suffer from any bias, the sampling distribution of the covariates holds no information about the covariate effects. It is therefore natural to exclude this uninformative distribution from the likelihood and carry out a conditional analysis which considers the data as coming from independent but non-identically distributed random variables given the covariates, though the sampled observations come from the joint distribution of the variable of interest and the covariates. This conditional point of view allows for a reduction of the dimensionality of the problem at hand. In the context of prevalent cohorts, as the observed data are sampled with a bias, the covariates are also not representative of the population of interest. The problem thereby extends beyond the length bias category. Some of the related sampling issues were discussed by Patil and Rao (1978), and Patil, Rao and Zelen (1988), with examples relating to family size estimation and zoology. Sackett (1979) offers an extensive and detailed catalog of the different kinds of bias that can occur in analytical research. Smith (1993) discussed the difficulties of defining target populations and the role of selection in inference. On a lighter note, Breslow (2003) discussed selection bias in epidemiology and the apparent lack of recognition of statistics in medicine.

The development of techniques taking into account biased sampling in specific problems and the underlying theory has grown through the years. Bickel and Ritov (1991) studied biased sampling and regression models when covariates have known finite support and offered large sample asymptotics for nonparametric linear regression. Zhou et al. (2002) proposed semiparametric empirical likelihood methods for outcome dependent sampling where they treat the distribution of the covariates as a nuisance parameter. Other works relating to outcome dependent sampling include Breslow and Holubkov (1997), Breslow, McNeney and Wellner (2003). Begg and Greenes (1983) studied selection bias and diagnostic tests, while Glesby and Hoover (1996) addressed the issue of treatment selection bias in observational studies. The use of biased samples to estimate treatment effect was investigated by Robbins and Zhang (1988), and estimation of incidence rates under biased sampling was studied by Berger, Bodian and Hirsch (1996). Lawless, Kalbfleisch and Wild (1999) proposed semiparametric methods for regression of problems with response-selective and missing data, and Chen (2001) introduced parametric models for response biased sampling. Recently, Baker, Fitzmaurice, Freedman and Kramer (2006) investigated informative covariates giving rise to missing outcomes.

As far as we know, very few authors have specifically addressed the sampling bias in the covariates with length-biased data. It is the case of Begg and Gray (1987) for case-control studies with prevalent cases, with the goal of estimating odds ratio relating exposure to disease incidence which is a very different problem from that of modeling the survival distribution. Cristóbal and Alcalá (2000) and Cristobál et al (2004) who proposed methods of nonparametric regression for length-biased observations based on a moment approach, also include the sampling bias in the covariates. They did not, however, consider censoring and therefore the remarks about Wang's approach applies to them as well.

When the stationarity assumption holds, it is possible to formalize the relationship between length-bias in the lifetimes and the bias in the covariates to show that the sampling distribution of the covariates depend on their effects. This then allows us to incorporate the extra information in the sampling distribution of the covariates and arrive at more efficient estimates.

Chapter 3

The likelihood function

In this chapter the likelihood function for length-biased data taken from prevalent cohort with right censoring including covariates is constructed, first through an extension of the likelihood of Vardi (1989) then it is formally derived from first principles. A comparison between this likelihood and the likelihood that ignores the contribution of the sampling distribution of the covariates is offered, both theoretically and through one analytic example.

3.1 Notation and preliminaries

3.1.1 Length-biased sampling

Let Y be a positive random variable (true event time) with distribution F_U.

Define, for $y \geq 0$, the c.d.f.

$$F_{LB}(y) = \frac{1}{\mu} \int_0^y x dF_U(x), \tag{3.1.1}$$

where $\mu = \int_0^\infty x dF_U(x) < \infty$.

The c.d.f. F_{LB} is called the length-biased distribution function of F_U and arises if a r.v. Y, with c.d.f. F_U is observed with probability proportional to its length (Cox (1969)). Let $\tilde{Y}$ have c.d.f. F_{LB}. The c.d.f. F_{LB} can also be described as the c.d.f of the randomly left-truncated r.v.'s $\tilde{Y}$ in the stationary case described by Wang (1991).

Now consider a sample $\tilde{Y}_1, \ldots, \tilde{Y}_n$ of independent random variables with c.d.f F_{LB}. In the applications considered, the $\tilde{Y}_i$'s are of the form $\tilde{Y}_i = T_i + R_i$, where T_i is the truncation variable (backward recurrence time) and R_i the observed residual lifetime (forward recurrence time). These correspond, respectively to the observed "onset" to the date of the cross sectional survey, and the observed time from the latter date until "failure". Suppose that the residual lifetimes are subject to censoring by the random variables C_i, with c.d.f $F_C(c)$, so that only the minimum of R_i and C_i is observed. Then the observed and possibly censored data are $\{(X_i, \delta_i), i = 1, 2, \cdots, n\}$ where $X_i = T_i + R_i \wedge C_i$ and where

$$\delta_i = \begin{cases} 1, & \text{if } \; R_i \leq C_i \\ 0, & \text{if } \; R_i > C_i \end{cases} . \tag{3.1.2}$$

These kinds of data can arise in survival analysis from sampling of prevalent cases, where subjects are selected at some point in time after onset and then followed until failure or censorship. Censoring may occur at the end of the follow-up period but is not restricted to this scenario. As mentioned in the previous chapter, for the observed lifetimes from a prevalent cohort to be properly length-biased, an important assumption is that, prior to the start of the study, the occurrences of initiations are assumed to follow a stationary Poisson process (Asgharian et al. 2006). A similar setting can occur in the sampling of renewal processes, when the process is observed from a starting time in between renewal events.

3.1.2 Introducing covariates

Now consider a cross-sectional survey, where the lifetimes are measured and, for each observation, a vector of covariates of interest, $\mathbf{Z}_i$, is assessed. The goal is to estimate F_U, by observing the data $\{(T_i, R_i \wedge C_i, \mathbf{Z}_i, \delta_i), i = 1, \ldots, n\}$, including the estimation of covariate effects on survival from initiation. Under this setting, the lifetimes given $\mathbf{Z}_i$ are not marginally identically distributed, but the samples come from a joint population of the lifetimes and covariates, having sampling distribution $F_{LB}(x, \mathbf{z})$ coming from the unbiased joint distribution $F_U(x, \mathbf{z})$ with parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\psi})$. Here, $\boldsymbol{\beta}$ encompasses the covariate effects and $\boldsymbol{\psi}$ represents the parameters of the "baseline" lifetime distribution, that is, the unbiased distribution of lifetime from onset when all covariates are set to 0.

In the usual, unbiased sampling case, the marginal sampling distribution of the covariates does not depend on the parameters of interest (in particular, $\boldsymbol{\beta}$), and one can carry out a conditional analysis to model the effect of the covariates on lifetimes, though the observations, sampled cross-sectionally, come from the joint distribution of the lifetime and covariates. When the covariates have an influence on the lifetime and the observations are sampled with length-bias, then the sampling distribution of the covariates is also subject to a bias and differs from the marginal distribution of the covariates in the (unbiased) population of interest, say $F_{\mathbf{Z}}(\mathbf{z})$. The marginal sampling distribution of the covariates, noted $F_B(\mathbf{z})$ will in fact be subject to a "mean lifetime" bias (as it will be shown in the next section), in the sense that the probability of observing the vector $\mathbf{z}$ will be proportional to the mean lifetime of an individual with covariates $\mathbf{z}$, and thus $F_B(\mathbf{z})$ will depend on $\boldsymbol{\theta}$. Consequently, the sampling distribution of the covariates has to be taken into account in the likelihood for inferential purposes.

3.2 From a conditional to a joint approach

Deriving the likelihood for length-biased observations subject to censoring in the presence of covariates needs careful consideration. While the issue of informative censoring (given the complete failure or censoring time from onset) can be dealt with using the approach of Vardi (1989), one cannot naively extend this likelihood to a likelihood that considers the covariates conditionally, because the distribution of the covariates in the sample also depends on the parameters of interest.

Specifically, consider the observed data as $W_i = (T_i, R_i \wedge C_i, \delta_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, rather than the conventional $(X_i, \delta_i, \mathbf{Z}_i)$, where $X_i = T_i + R_i \wedge C_i$. While complete lifetimes and complete censoring times are *not* independent (because of the common backward recurrence time), the residual lifetimes R_i and residual censoring times C_i are, in most practical situations. Further, we shall assume that the residual censoring times are independent of the (T_i, R_i) pairs, $i = 1, \ldots, n$, and so the W_i's are assumed to be independent.

In the absence of covariates, and under a scheme of multiplicative censoring, which is informative, Vardi derives the likelihood as,

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}), \tag{3.2.1}$$

where

$$\mathcal{L}_i(\boldsymbol{\theta}) = \Big(\frac{f_U(u_i; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\Big)^{\delta_i} \Big(\int_{w \geq v_i} \frac{f_U(w; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\, dw\Big)^{1-\delta_i}, \tag{3.2.2}$$

and f_U is the density function of F_U, while $\mu(\boldsymbol{\theta})$ the mean of f_U. A detailed description of Vardi's problem and the derivation of 3.2.1 is given in section 6.1. Under cross-sectional sampling, the likelihood is proportional to $\mathcal{L}(\boldsymbol{\theta})$, though the setup differs from multiplicative censoring, as Vardi explicitly states. The asymptotic properties of MLE's obtained from $\mathcal{L}(\boldsymbol{\theta})$ under cross-sectional sampling without covariates are derived by Asgharian et al (2002a). When covariates are introduced in the model, the following likelihood:

$$\mathcal{L}_I(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left(\frac{f_U(u_i|\mathbf{z}_i;\boldsymbol{\theta})}{\mu(\mathbf{z}_i;\boldsymbol{\theta})} \right)^{\delta_i} \left(\int_{w \geq v_i} \frac{f_U(w|\mathbf{z}_i;\boldsymbol{\theta})}{\mu(\mathbf{z}_i;\boldsymbol{\theta})}\, dw \right)^{1-\delta_i} \tag{3.2.3}$$

where $\mu(\mathbf{z}_i;\boldsymbol{\theta}) = \mathbb{E}(X|\mathbf{z}_i)$, seems natural, as it conditions over the observed covariates as is done in most common regression contexts. This likelihood, however, *ignores* the informativeness of the sampling distribution of the covariates (hence the subscript I in $\mathcal{L}_I$).

Going back to first principles to properly include covariates, in the most general setting, the likelihood can be seen as the joint density of the data. Let f denote a generic joint density, by independence of the observations:

$$\mathcal{L}_J(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i, t_i, \mathbf{z}_i | X \geq T; \boldsymbol{\theta}). \tag{3.2.4}$$

Here, the subscript J is used to denote the *joint* approach, as the sampling distribution of the covariates is not conditioned out. The condition $X \geq T$ takes into account the fact that every observation is randomly truncated, and essentially reflects the sampling mechanism.

There are three approaches one can take to go from 3.2.4 to something with a form similar to 3.2.2 (in fact the former should reduce to the latter when there are no covariates). All three approaches lead to the same likelihood, though the formula may be written in different ways. These aesthetic adjustments depend on the intended emphasis: for optimization purposes, all parts not involving $\boldsymbol{\theta}$ may be ignored, while theoretical results are easier to perceive through a fully joint likelihood.

The first approach consists of applying the definition of conditional probability in its most basic form, (looking at just one observation for simplicity):

$$\mathcal{L}_J(\boldsymbol{\theta}) = f(x, t, \mathbf{z}|X \geq T; \boldsymbol{\theta}) = \frac{f(x, t, \mathbf{z}; \boldsymbol{\theta})}{P(X \geq T; \boldsymbol{\theta})} = \frac{f(x, t|\mathbf{z}; \boldsymbol{\theta}) f_{\mathbf{Z}}(\mathbf{z})}{P(X \geq T; \boldsymbol{\theta})} \tag{3.2.5}$$

As $f_{\mathbf{Z}}(\mathbf{z})$ does not depend on $\boldsymbol{\theta}$, this reduces to:

$$\mathcal{L}_J(\boldsymbol{\theta}) \propto \frac{f(x|\mathbf{z}; \boldsymbol{\theta}) f(t|\mathbf{z}; \boldsymbol{\theta})}{P(X \geq T; \boldsymbol{\theta})}. \tag{3.2.6}$$

We can assume that the truncation time is independent of the covariates, so $f(t|\mathbf{z}; \boldsymbol{\theta}) = f_T(t; \boldsymbol{\theta})$. In fact, the assumption of stationary incidence of onset times implies uniform truncation which in turn is equivalent to length-bias sampling, so $f_T(t; \boldsymbol{\theta})$ is proportional to a constant. Thus,

$$\mathcal{L}_J(\boldsymbol{\theta}) \propto \frac{f(x|\mathbf{z}; \boldsymbol{\theta})}{P(X \geq T; \boldsymbol{\theta})}. \tag{3.2.7}$$

The second approach uses the relationship between joint and conditional density (once again, considering only one observation for simplicity):

$$\mathcal{L}_J(\boldsymbol{\theta}) = f(x, t, \mathbf{z}|X \geq T; \boldsymbol{\theta}) = f(x, t|\mathbf{z}, X \geq T; \boldsymbol{\theta}) f(\mathbf{z}|X \geq T; \boldsymbol{\theta}). \tag{3.2.8}$$

This approach enables us to see that

$$\mathcal{L}_J(\boldsymbol{\theta}) = \mathcal{L}_I(\boldsymbol{\theta}) f_B(\mathbf{z};\boldsymbol{\theta}). \tag{3.2.9}$$

So let us look at $f_B(\mathbf{z};\boldsymbol{\theta})$, the all important distinction between $\mathcal{L}_I(\boldsymbol{\theta})$ and $\mathcal{L}_J(\boldsymbol{\theta})$:

$$f_B(\mathbf{z}|\boldsymbol{\theta}) = f(\mathbf{z}|X \geq T;\boldsymbol{\theta}) = \frac{P(X \geq T|\mathbf{z};\boldsymbol{\theta}) f_{\mathbf{Z}}(\mathbf{z})}{P(X \geq T;\boldsymbol{\theta})} \tag{3.2.10}$$

So we need to find $P(X \geq T|\mathbf{z};\boldsymbol{\theta})$.

$$P(X \geq T|\mathbf{z};\boldsymbol{\theta}) = \int_0^\infty \int_0^x f_U(x,t|\mathbf{z};\boldsymbol{\theta})\, dt\, dx \tag{3.2.11}$$

Since the unbiased random variables X and T are independent, and, as stated above, the truncation is assumed to be uniform (and independent of the covariates), we get:

$$P(X \geq T|\mathbf{z};\boldsymbol{\theta}) = \int_0^\infty \int_0^x f_U(x|\mathbf{z};\boldsymbol{\theta}) f_T(t)\, dt\, dx \tag{3.2.12}$$

Let τ be some large constant, covering the range of possible lifetimes in the population of interest (as we want the truncation to have a proper distribution, though the distribution of the lifetimes might have in theory unbounded support on the positive real line).

$$P(X \geq T|\mathbf{z};\boldsymbol{\theta}) = \int_0^\infty \int_0^x f_U(x|\mathbf{z};\boldsymbol{\theta})\frac{1}{\tau}\,dt\,dx = \int_0^\infty \frac{1}{\tau} x f_U(x|\mathbf{z};\boldsymbol{\theta})\,dx = \frac{\mu(\mathbf{z};\boldsymbol{\theta})}{\tau} \tag{3.2.13}$$

Hence,

$$f_B(\mathbf{z};\boldsymbol{\theta}) = f(\mathbf{z}|X \geq T;\boldsymbol{\theta}) = \frac{P(X \geq T|\mathbf{z};\boldsymbol{\theta}) f_{\mathbf{Z}}(\mathbf{z})}{\int_{\mathbf{z}} P(X \geq T|\mathbf{z}) f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z}} \tag{3.2.14}$$

$$f_B(\mathbf{z};\boldsymbol{\theta}) = \frac{\frac{1}{\tau}\mu(\mathbf{z};\boldsymbol{\theta}) f_{\mathbf{Z}}(\mathbf{z})}{\frac{1}{\tau}\int_{\mathbf{z}} \mu(\mathbf{z};\boldsymbol{\theta}) f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z}} = \frac{\mu(\mathbf{z};\boldsymbol{\theta}) f_{\mathbf{Z}}(\mathbf{z})}{\mu(\boldsymbol{\theta})}, \tag{3.2.15}$$

where $\mu(\boldsymbol{\theta}) = \mathbb{E}(\mathbb{E}(X|\mathbf{Z})) = \mathbb{E}(X)$, the overall mean lifetime of the unbiased population. Note that, as mentioned earlier, the covariates are sampled proportionally to the mean lifetime of a given set of covariates. Substituting 3.2.15 back into 3.2.9, one gets:

$$\mathcal{L}_J(\boldsymbol{\theta}) \propto \mathcal{L}_I(\boldsymbol{\theta}) \frac{\mu(\mathbf{z};\boldsymbol{\theta}) f_{\mathbf{Z}}(\mathbf{z})}{\mu(\boldsymbol{\theta})} \tag{3.2.16}$$

Considering a full sample the likelihood is thus:

$$\mathcal{L}_J(\boldsymbol{\theta}) \propto \mathcal{L}_I(thev) \left(\prod_{i=1}^{n} \frac{\mu(\mathbf{z}_i;\boldsymbol{\theta}) f_{\mathbf{Z}}(\mathbf{z}_i)}{\mu(\boldsymbol{\theta})} \right) \tag{3.2.17}$$

After some algebraic manipulation, eliminating all terms not depending on

$\boldsymbol{\theta}$ and canceling the conditional expectations, yields:

$$\mathcal{L}_J(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \left(\frac{f_U(u_i|\mathbf{z}_i;\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} \right)^{\delta_i} \left(\int_{w \geq v_i} \frac{f_U(w|\mathbf{z}_i;\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} \, dw \right)^{1-\delta_i} \tag{3.2.18}$$

As one can notice in 3.2.15, $P(X \geq T)$ is proportional to $\mu(\boldsymbol{\theta})$, and therefore 3.2.7 also gives rise to 3.2.18.

The third approach is a natural extension of 3.2.2 into a sample space $\mathcal{X}$ of higher dimension. Here we keep $(X, \mathbf{Z})$ as joint, though marginally the unbiased distribution of $\mathbf{Z}$ holds no information about $\boldsymbol{\theta}$, working in a higher dimension simplifies the proofs seen in the next chapter. We have IID observations of length-biased times accompanied by covariates, so $\mu(\boldsymbol{\theta})$ is indeed the mean lifetime of the entire population with distribution $F_U(x, \mathbf{z})$. Under this approach, we obtain:

$$\mathcal{L}_J \propto \prod_{i=1}^{n} \left(\frac{f_U(u_i, \mathbf{z}_i;\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} \right)^{\delta_i} \left(\int_{w \geq v_i} \frac{f_U(w, \mathbf{z}_i;\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} \, dw \right)^{1-\delta_i} \tag{3.2.19}$$

3.3 Formal derivation of the likelihood function

Following is the actual derivation of the likelihood function.

Let $B_{d\mathbf{z}_i}(\mathbf{z}_i)$ be a ball of radius $d\mathbf{z}_i$ centered at $\mathbf{z}_i$.

$$\begin{aligned} P(X > x) &= \int_{\mathbf{z}} P(X > x, \mathbf{Z} = \mathbf{z})\, d\mathbf{z} = \int_{\mathbf{z}} P(T + R \wedge C > x, \mathbf{Z} = \mathbf{z})\, d\mathbf{z} \\ &= \int_{\mathbf{z}} \int_0^{\infty} P(T + R \wedge C > x, \mathbf{Z} = \mathbf{z} | C = c) f_C(c)\, dc\, d\mathbf{z} \\ &= \int_{\mathbf{z}} \int_0^{\infty} P(T + R \wedge c > x, \mathbf{Z} = \mathbf{z}) f_C(c)\, dc\, d\mathbf{z}. \end{aligned} \tag{3.3.1}$$

and

$$\begin{aligned} P(T + R \wedge c > x, \mathbf{Z} = \mathbf{z}) &= \int_r P(T + R \wedge c > x, \mathbf{Z} = \mathbf{z}, R = r)\, dr \\ &= \int_r \int_{t > x - (r \wedge c)} f_{T,R,\mathbf{Z}}(t, r, \mathbf{z})\, dt\, dr. \end{aligned} \tag{3.3.2}$$

Thus,

$$\begin{aligned} S_X(x) &= \int_{\mathbf{z}} \int_c \int_r \int_{t > x - (r \wedge c)} f_{T,R,\mathbf{z}}(t, r, \mathbf{z}) f_C(c)\, dt\, dr\, dc\, d\mathbf{z} \\ &= \int_c \int_{\mathbf{z}} \int_r \int_{t > x - (r \wedge c)} f_{T,R,\mathbf{z}}(t, r, \mathbf{z})\, dt\, dr\, d\mathbf{z} f_C(c)\, dc \\ &= \int_c \left[\int_{\mathbf{z}} \int_r \int_{t > x - (r \wedge c)} \frac{f_U(t + r, \mathbf{z})}{\mu}\, dt\, dr\, d\mathbf{z} \right] f_C(c)\, dc \\ &= \mathbb{E}_C \big[\Phi_x(C) \big], \end{aligned} \tag{3.3.3}$$

where

$$\Phi_x(c) = \int_{\mathbf{z}} \int_r \int_{t > x - r \wedge c} \frac{f_U(a + r, \mathbf{z})}{\mu}\, dt\, dr\, d\mathbf{z}. \tag{3.3.4}$$

The distribution of $T + C$, the censoring time, is

$$f_{T+C}(\omega) = \int_{\mathbf{z}} \int_0^{\omega} f_{T,\mathbf{Z}}(\omega - c, \mathbf{z}) f_C(c)\, dc\, d\mathbf{z} = (f_T * f_C)(\omega), \qquad (3.3.5)$$

where $*$ is the convolution operator.

The likelihood function is

$$
\begin{aligned}
\mathcal{L} = \prod_{i=1}^{n} & P^{\delta_i}\big(T_i \in (t_i, t_i + dt_i], R_i \in (r_i, r_i + dr_i], \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) \times \\
& \qquad P^{1-\delta_i}\big(T_i \in (t_i, t_i + dt_i], C_i \in (c_i, c_i + dc_i], \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) \\
= \prod_{i \in UC} & P\big(T_i \in (t_i, t_i + dt_i], R_i \in (r_i, r_i + dr_i], R_i \le C_i, \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) \times \\
& \prod_{i \in C} P\big(T_i \in (t_i, t_i + dt_i], C_i \in (c_i, c_i + dc_i], R_i > C_i, \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) \\
= \prod_{i \in UC} & P\big(R_i \le C_i \mid T_i \in (t_i, t_i + dt_i], R_i \in (r_i, r_i + dr_i], \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) \times \\
& \qquad P\big(T_i \in (t_i, t_i + dt_i], R_i \in (r_i, r_i + dr_i], \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) \times \\
& \prod_{i \in C} P\big(R_i > C_i \mid T_i \in (t_i, t_i + dt_i], C_i \in (c_i, c_i + dc_i], \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) \times \\
& \qquad P\big(T_i \in (t_i, t_i + dt_i], C_i \in (c_i, c_i + dc_i], \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) \\
= \prod_{i \in UC} & P\big(C_i \ge r_i \mid T_i \in (t_i, t_i + dt_i], R_i \in (r_i, r_i + dr_i], \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) \times \\
& \qquad P\big(T_i \in (t_i, t_i + dt_i], R_i \in (r_i, r_i + dr_i], \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) \times \\
& \prod_{i \in C} P\big(R_i > c_i \mid T_i \in (t_i, t_i + dt_i], C_i \in (c_i, c_i + dc_i], \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) \times \\
& \qquad P\big(T_i \in (t_i, t_i + dt_i], \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) P\big(C_i \in (c_i, c_i + dc_i]\big) \\
= \prod_{i \in UC} & P\big(C_i \ge r_i\big) P\big(T_i \in (t_i, dt_i], R_i \in (r_i, dr_i], \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) \times \\
& \prod_{i \in C} P\big(R_i > c_i \mid T_i \in (t_i, dt_i], \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) P\big(T_i \in (t_i, dt_i], \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) \times \\
& \quad P\big(C_i \in (c_i, dc_i]\big)
\end{aligned}
\tag{3.3.6}
$$

As C does not contain any information about $\boldsymbol{\theta}$, we obtain

$$
\begin{aligned}
\mathcal{L} \propto & \prod_{i \in UC} P\big(T_i \in (t_i, dt_i], R_i \in (r_i, dr_i], \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) \times \\
& \prod_{i \in C} P\big(R_i > c_i, T_i \in (t_i, dt_i], \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) \\
= & \prod_{i \in UC} f_{T,R,\mathbf{Z}}(t_i, r_i, \mathbf{z}_i) dt_i dr_i d\mathbf{z}_i \prod_{i \in C} P_{T,R}\big(R > c_i, T \in (t_i, dt_i], \mathbf{Z}_i \in B_{d\mathbf{z}_i}(\mathbf{z}_i)\big) \\
\propto & \prod_{i \in UC} f_{T,R,\mathbf{Z}}(t_i, r_i, \mathbf{z}_i) \prod_{i \in C} \int_{c_i}^{\infty} f_{T,R,\mathbf{Z}}(t_i, r, \mathbf{z}) dr \\
= & \prod_{i=1}^{n} f_{T,R,\mathbf{Z}}^{\delta_i}(t_i, r_i, \mathbf{z}_i) \Big(\int_{c_i}^{\infty} f_{T,R,\mathbf{Z}}(t_i, r, \mathbf{z}) dr \Big)^{1-\delta_i}.
\end{aligned}
\tag{3.3.7}
$$

Using the fact that

$$
f_{T,R,\mathbf{Z}}(t, r, \mathbf{z}) = \frac{f_U(t + r, \mathbf{z})}{\mu} \quad \text{if } a > 0, \; r > 0. \tag{3.3.8}
$$

we obtain

$$
\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} \Big(\frac{f_U(t_i + r_i, \mathbf{z}_i)}{\mu} \Big)^{\delta_i} \Big(\int_{w \geq c_i} \frac{f_U(t_i + w, \mathbf{z}_i)}{\mu} dw \Big)^{1-\delta_i} \\
&= \prod_{i=1}^{n} \Big(\frac{f_U(t_i + r_i, \mathbf{z}_i)}{\mu} \Big)^{\delta_i} \Big(\int_{\omega \geq t_i + c_i} \frac{f_U(\omega, \mathbf{z}_i)}{\mu} d\omega \Big)^{1-\delta_i}.
\end{aligned}
\tag{3.3.9}
$$

Let $u_i = t_i + r_i$ and $v_i = t_i + c_i$, then

$$
\mathcal{L} = \prod_{i=1}^{n} \Big(\frac{f_U(u_i, \mathbf{z}_i)}{\mu} \Big)^{\delta_i} \Big(\int_{\omega \geq v_i} \frac{f_U(\omega, \mathbf{z}_i)}{\mu} d\omega \Big)^{1-\delta_i}. \tag{3.3.10}
$$

In a parametric setting,

$$\mathcal{L}_i(\boldsymbol{\theta}) = \Big(\frac{f_U(u_i, \mathbf{z}_i; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\Big)^{\delta_i} \Big(\int_{\omega \geq v_i} \frac{f_U(\omega, \mathbf{z}_i; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\, d\omega\Big)^{1-\delta_i} \tag{3.3.11}$$

and

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}), \tag{3.3.12}$$

3.4 Relationship between $\mathcal{L}_I(\boldsymbol{\theta})$ and $\mathcal{L}_J(\boldsymbol{\theta})$.

It is worthwhile to compare the likelihood that ignores the sampling distribution of the covariates and this new joint biased-sampling likelihood. One may notice that 3.2.18 differs from 3.2.7 only in the replacement of $\mu(\mathbf{z}_i; \boldsymbol{\theta})$ by $\mu(\boldsymbol{\theta})$. By definition,

$$\mu(\boldsymbol{\theta}) = \int_{\mathbf{z}} \mu(\mathbf{z}; \boldsymbol{\theta}) f_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z}, \tag{3.4.1}$$

so we go from n conditional expectations to one unconditional mean, but this does not say much about the context of biased sampling. A more intriguing result arises from 3.2.15:

$$\frac{f_{\mathbf{Z}}(\mathbf{z})}{\mu(\boldsymbol{\theta})} = \frac{f_B(\mathbf{z}; \boldsymbol{\theta})}{\mu(\mathbf{z}; \boldsymbol{\theta})}. \tag{3.4.2}$$

Integrating over $\mathbf{z}$ on each side, one gets:

$$\frac{1}{\mu(\boldsymbol{\theta})} = \int_{\mathbf{z}} \frac{f_B(\mathbf{z};\boldsymbol{\theta})}{\mu(\mathbf{z};\boldsymbol{\theta})}\, d\mathbf{z} = \mathbb{E}_B\left(\frac{1}{\mu(\mathbf{Z};\boldsymbol{\theta})}\right), \tag{3.4.3}$$

and so we have the relation between $\mu(\boldsymbol{\theta})$ and $\mu(\mathbf{z};\boldsymbol{\theta})$ in the presence of "expected-lifetime" bias in the sampling distribution of the covariates. Equation 3.4.3 suggests that

$$\hat{\mu}(\boldsymbol{\theta}) = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\mu(\mathbf{z}_i;\boldsymbol{\theta})}\right)^{-1} \tag{3.4.4}$$

is a natural estimate for $\mu(\boldsymbol{\theta})$. Replacing $\mu(\boldsymbol{\theta})$ by $\hat{\mu}(\boldsymbol{\theta})$ in 3.2.19 one can obtain a tractable profile likelihood. In Chapter 5, we discuss why it can be difficult to use this estimate for $\mu(\boldsymbol{\theta})$ to find MLEs.

In the interest of implementation and asymptotics, we can now compare $\mathcal{L}_I(\boldsymbol{\theta})$ and $\mathcal{L}_J(\boldsymbol{\theta})$. Since it is easier to work on the log scale, one can write:

$$\ell_I(\boldsymbol{\theta}) = \ell_\cap - \sum_{i=1}^{n} \log\left(\mu(\mathbf{z}_i;\boldsymbol{\theta})\right), \tag{3.4.5}$$

$$\ell_J(\boldsymbol{\theta}) = \ell_\cap - n\log\left(\mu(\boldsymbol{\theta})\right), \tag{3.4.6}$$

where the common part of the two log-likelihoods is given by

$$\ell_\cap = \sum_{i=1}^{n} \delta_i \log\left(f_U(x_i|\mathbf{z}_i;\boldsymbol{\theta})\right) + \sum_{i=1}^{n} (1-\delta_i)\log\left(\int_{w\geq x_i} f_U(w|\mathbf{z}_i;\boldsymbol{\theta})\, dw\right). \tag{3.4.7}$$

Now, considering the biased sampling of $\mathbf{Z}$,

$$-\log\left(\mu(\boldsymbol{\theta})\right) = \log\left(\frac{1}{\mu(\boldsymbol{\theta})}\right) = \log\left(\mathbb{E}_B\left(\frac{1}{\mu(\mathbf{Z};\boldsymbol{\theta})}\right)\right). \tag{3.4.8}$$

On the other hand, considering the Strong Law of Large Numbers,

$$-\frac{1}{n}\sum_{i=1}^{n}\log\left(\mu(\mathbf{z}_i;\boldsymbol{\theta})\right) = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{\mu(\mathbf{z}_i;\boldsymbol{\theta})}\right) \xrightarrow{a.s.} \mathbb{E}_B\left(\log\left(\frac{1}{\mu(\mathbf{Z};\boldsymbol{\theta})}\right)\right). \tag{3.4.9}$$

Using Jensen's inequality, by concavity of the logarithm, we can see that for large enough samples

$$\ell_J(\boldsymbol{\theta}) \geq \ell_I(\boldsymbol{\theta}) \Longrightarrow \mathcal{L}_J(\boldsymbol{\theta}) \geq \mathcal{L}_I(\boldsymbol{\theta}) \tag{3.4.10}$$

The last question that comes to mind with respect to the comparison between the two likelihoods is what about the maximum likelihood estimates? In particular, how would $\hat{\boldsymbol{\theta}}_{I,n}$ and $\hat{\boldsymbol{\theta}}_{J,n}$ compare in matters of bias and variance?

Theorem 1. *Suppose $\boldsymbol{\Theta}$ is a bounded, open subset of $\mathbb{R}^k$. Let $f_U(x|\mathbf{z};\boldsymbol{\theta})$ be identifiable on $\overline{\boldsymbol{\Theta}}$, the closure of $\boldsymbol{\Theta}$. Let $\mu(\mathbf{z};\boldsymbol{\theta}) : \boldsymbol{\Theta} \to (0,+\infty)$ be a differentiable function of $\boldsymbol{\theta}$ and suppose there exists an integrable function $m(\mathbf{z},\boldsymbol{\theta})$ such that $|\frac{\partial}{\partial\boldsymbol{\theta}}\mu(\mathbf{z},\boldsymbol{\theta})| \leq m(\mathbf{z},\boldsymbol{\theta})$. Then, under sampling from the* *joint biased distribution of* $(X, \mathbf{Z})$

$$\left| \frac{\partial}{\partial \boldsymbol{\theta}} \ell_{I,n}(\boldsymbol{\theta}) - \frac{\partial}{\partial \boldsymbol{\theta}} \ell_{J,n}(\boldsymbol{\theta}) \right| \xrightarrow{a.s.} 0 \tag{3.4.11}$$

Proof. As mentioned earlier in this chapter, both log-likelihoods can be expressed in terms of a common log-likelihood plus the sum of n log mean lifetimes. Looking at their gradients yields (here we use the $'$ notation as short-hand for first derivative with respect to $\boldsymbol{\theta}$):

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell_{I,n}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\cap,n}(\boldsymbol{\theta}) - \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log \mu(\mathbf{z}_i; \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\cap,n}(\boldsymbol{\theta}) - \sum_{i=1}^{n} \frac{\mu'(\mathbf{z}_i; \boldsymbol{\theta})}{\mu(\mathbf{z}_i; \boldsymbol{\theta})}, \tag{3.4.12}$$

and

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell_{J,n}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\cap,n}(\boldsymbol{\theta}) - n \frac{\partial}{\partial \boldsymbol{\theta}} \log \mu(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\cap,n}(\boldsymbol{\theta}) - n \frac{1}{\mu(\boldsymbol{\theta})} \frac{\partial \mu(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{3.4.13}$$

Now,

$$\frac{\partial \mu(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} \int \mu(\mathbf{z}; \boldsymbol{\theta}) f_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z}, \tag{3.4.14}$$

At this point it becomes imperative to interchange integration and differentiation. In order to do so, we imposed a smoothness condition on $\mu(\mathbf{z}, \boldsymbol{\theta})$ through the existence of an integrable function $m(\mathbf{z}, \boldsymbol{\theta})$ which dominates $\mu(\mathbf{z}, \boldsymbol{\theta})$. See Folland (1999), Theorem 2.27 page 56 for details.

$$\frac{\partial \mu(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \int \frac{\partial}{\partial \boldsymbol{\theta}} \mu(\mathbf{z}, \boldsymbol{\theta}) f_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z} = \int \mu'(\mathbf{z}; \boldsymbol{\theta}) f_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z}. \tag{3.4.15}$$

Note that, by 3.2.15,

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{\mu(\boldsymbol{\theta})}{\mu(\mathbf{z};\boldsymbol{\theta})} f_B(\mathbf{z};\boldsymbol{\theta}), \tag{3.4.16}$$

therefore

$$\frac{\partial \mu}{\partial \boldsymbol{\theta}} = \int \mu'(\mathbf{z};\boldsymbol{\theta}) \frac{\mu(\boldsymbol{\theta})}{\mu(\mathbf{z},\boldsymbol{\theta})} f_B(\mathbf{z};\boldsymbol{\theta})\, d\mathbf{z} = \mu(\boldsymbol{\theta}) \mathbb{E}_B \left[\frac{\mu'(\mathbf{Z};\boldsymbol{\theta})}{\mu(\mathbf{Z};\boldsymbol{\theta})} \right]. \tag{3.4.17}$$

So

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell_{J,n}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\cap,n}(\boldsymbol{\theta}) - n \mathbb{E}_B \left[\frac{\mu'(\mathbf{Z};\boldsymbol{\theta})}{\mu(\mathbf{Z};\boldsymbol{\theta})} \right] \tag{3.4.18}$$

and

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell_{I,n}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\cap,n}(\boldsymbol{\theta}) - n \frac{1}{n} \sum_{i=1}^{n} \frac{\mu'(\mathbf{z}_i;\boldsymbol{\theta})}{\mu(\mathbf{z}_i;\boldsymbol{\theta})} \tag{3.4.19}$$

When sampling from the biased distribution $\mathbf{Z}$, the strong law of large numbers gives us that

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\mu'(\mathbf{z}_i;\boldsymbol{\theta})}{\mu(\mathbf{z}_i;\boldsymbol{\theta})} \overset{a.s.}{\to} \mathbb{E}_B \left[\frac{\mu'(\mathbf{Z};\boldsymbol{\theta})}{\mu(\mathbf{Z};\boldsymbol{\theta})} \right], \tag{3.4.20}$$

hence $|\frac{\partial}{\partial \boldsymbol{\theta}} \ell_{I,n}(\boldsymbol{\theta}) - \frac{\partial}{\partial \boldsymbol{\theta}} \ell_{J,n}(\boldsymbol{\theta})| \overset{a.s.}{\to} 0.\Box$

Theorem 1 essentially states that the distance between the score functions of the likelihoods tends to 0. Given identifiability of the parameters, the asymptotics of $\mathcal{L}_I(\boldsymbol{\theta})$ as derived by Asgharian et al. (2002a) and the asymptotic properties of $\mathcal{L}_J(\boldsymbol{\theta})$ we derive in Chapter 4 of this thesis, this means that both likelihoods have the same maximizers, and therefore will give the same point estimates for $\boldsymbol{\theta}$, as the sample size n tends to infinity. The issue of efficiency is more subtle. It is shown in the next section with an analytic example and by simulation in Chapter 6 that one can obtain more efficient estimates using $\mathcal{L}_J(\boldsymbol{\theta})$ when $F_{\mathbf{Z}}(\mathbf{z})$ is known.

3.5 Efficiency of $\hat{\boldsymbol{\theta}}_J$ over $\hat{\boldsymbol{\theta}}_I$: an analytic example

In Theorem 1, we showed that, at least asymptotically, maximization of $\mathcal{L}_I$ will result in the same point estimates for $\boldsymbol{\theta}$ as maximizing of $\mathcal{L}_J$ under sampling from the joint biased density of X and $\mathbf{Z}$. We now address the issue of efficiency by looking at the variances of $\hat{\boldsymbol{\theta}}_J$ and $\hat{\boldsymbol{\theta}}_I$ in a particular case, practically the simplest possible case for the regression of length-biased lifetimes. Consider exponentially distributed lifetimes (in the unbiased population) with a binary covariate that is known to have a discrete uniform distribution in the unbiased population. Suppose we have no censoring that is, $\delta = 1$ with probability 1. Using a proportional hazards model, the hazard function for the unbiased exponential population is given by:

$$\lambda(x|z) = e^{\beta z}\lambda. \tag{3.5.1}$$

Note that we essentially have a mixture of two exponentially distributed populations, one with rate $e^{\beta}\lambda$ and the other with rate λ.

With a little algebra, it can be shown that the length-biased density of an exponential random variable with rate λ is a Gamma density with shape parameter 2 and rate λ:

$$f_U(x;\lambda) = \lambda \exp(-\lambda x); \tag{3.5.2}$$

$$f_{LB}(x;\lambda) = \frac{x\lambda \exp(-\lambda x)}{\frac{1}{\lambda}} = \lambda^2 x \exp(-\lambda x) = \frac{\lambda^2 x^{2-1} \exp(-\lambda x)}{\Gamma(2)}. \tag{3.5.3}$$

For our model, this means the conditional length-biased density of the lifetimes, given z is:

$$f_{LB}(x|z) = (e^{\beta z}\lambda)^2 x \exp\Big(-e^{\beta z}\lambda x\Big) \tag{3.5.4}$$

Without loss of generality, let $Z \in \{0,1\}$. The sampling probability of Z is:

$$p_B(z) = \frac{e^{-\beta z}}{1+e^{-\beta}}. \tag{3.5.5}$$

The mean lifetime given z is

$$\mu(z;\beta,\lambda) = (e^{\beta z}\lambda)^{-1}, \tag{3.5.6}$$

end the overall unbiased mean lifetime is

$$\mu(\beta, \lambda) = \frac{1 + e^{-\beta}}{2\lambda}. \tag{3.5.7}$$

We consider the conditional likelihood first. As there is no censoring, the likelihood reduces to:

$$\begin{aligned}
\mathcal{L}_I(\mathbf{x}, \mathbf{z}|\beta, \lambda) &= \prod_{i=1}^{n} \frac{f_U(x_i|z_i)}{\mu(z; \beta, \lambda)} \\
&= \prod_{i=1}^{n} \frac{e^{\beta z_i}\lambda \exp(-e^{\beta z_i}\lambda x_i)}{(e^{\beta z_i}\lambda)^{-1}} \\
&= \prod_{i=1}^{n} e^{2\beta z_i}\lambda^2 \exp(-e^{\beta z_i}\lambda x_i).
\end{aligned} \tag{3.5.8}$$

The log-likelihood is given by:

$$\ell_I(\mathbf{x}, \mathbf{z}|\beta, \lambda) = \sum_{i=1}^{n} \Big(2\beta z_i + 2\log\lambda - e^{\beta z_i}\lambda x_i\Big). \tag{3.5.9}$$

The score function is then

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell_I(\boldsymbol{\theta}) = \begin{pmatrix} \sum_{i=1}^{n} \Big(2z_i - z_i e^{\beta z_i}\lambda x_i\Big) \\ \sum_{i=1}^{n} \Big(\frac{2}{\lambda} - e^{\beta z_i} x_i\Big) \end{pmatrix}. \tag{3.5.10}$$

At this point, it is relevant to look at the expectation of 3.5.10 under length-biased sampling, to verify that it is zero. Without loss of generality, we drop the subscript and consider only one observation. Three separate expectations need to be computed. First,

$$
\begin{aligned}
\mathbb{E}_{LB}(Z) &= \sum_{z=0}^{1} z p_B(z;\beta) \\
&= 1 \times \frac{e^{-\beta}}{1+e^{-\beta}} \\
&= \frac{1}{1+e^{\beta}}.
\end{aligned} \tag{3.5.11}
$$

Similarly,

$$
\begin{aligned}
\mathbb{E}_{LB}(Ze^{\beta Z}\lambda X) &= \int_0^\infty \sum_{z=0}^{1} z e^{\beta z}\lambda x f_{LB}(x|z;\beta,\lambda) p_B(z;\beta)\, dx \\
&= \frac{1}{1+e^{\beta}} \int_0^\infty e^{\beta}\lambda x f_{LB}(x|z=1;\beta,\lambda)\, dx \\
&= \frac{1}{1+e^{\beta}} \int_0^\infty e^{\beta}\lambda x (e^{\beta}\lambda)^2 x \exp(-e^{\beta}\lambda x)\, dx \\
&= \frac{1}{1+e^{\beta}} \Gamma(3) \int_0^\infty \frac{(e^{\beta}\lambda)^3}{\Gamma(3)} x^{3-1} \exp(-e^{\beta}\lambda x)\, dx \\
&= \frac{2}{1+e^{\beta}},
\end{aligned} \tag{3.5.12}
$$

by integrating out the density of a Gamma($3, e^{\beta}\lambda$) in the fourth line above.

The third expected value is slightly more tricky:

$$\begin{aligned}
\mathbb{E}_{LB}(e^{\beta Z}X) &= \int_0^\infty \sum_{z=0}^{1} e^{\beta z} x f_{LB}(x|z;\beta,\lambda) p_B(z;\beta)\, dx \\
&= \int_0^\infty x f_{LB}(x|0;\beta,\lambda)\frac{1}{1+e^{-\beta}}\, dx + \int_0^\infty e^{\beta} x f_{LB}(x|1;\beta,\lambda)\frac{e^{-\beta}}{1+e^{-\beta}}\, dx \\
&= I + II, \text{ say.}
\end{aligned} \tag{3.5.13}$$

Treating I and II separately:

$$\begin{aligned}
I &= \frac{1}{1+e^{-\beta}} \int_0^\infty x\lambda^2 x \exp(-\lambda x)\, dx \\
&= \frac{1}{1+e^{-\beta}} \mathbb{E}(Y), \text{ where } Y \sim \text{Gamma}(2,\lambda) \qquad (3.5.14) \\
&= \frac{1}{1+e^{-\beta}} \frac{2}{\lambda},
\end{aligned}$$

$$\begin{aligned}
II &= \frac{e^{-\beta}}{1+e^{-\beta}} \int_0^\infty e^{\beta} x (e^{\beta}\lambda)^2 x \exp(-e^{\beta}\lambda x)\, dx \\
&= \frac{1}{1+e^{-\beta}} \int_0^\infty x (e^{\beta}\lambda)^2 x \exp(-e^{\beta}\lambda x)\, dx \\
&= \frac{1}{1+e^{-\beta}} \mathbb{E}(Y), \text{ where } Y \sim \text{Gamma}(2, e^{\beta}\lambda) \qquad (3.5.15) \\
&= \frac{1}{1+e^{-\beta}} \frac{2}{e^{\beta}\lambda} \\
&= \frac{e^{-\beta}}{1+e^{-\beta}} \frac{2}{\lambda}.
\end{aligned}$$

Hence,

$$I + II = \frac{2}{\lambda}\left(\frac{1}{1+e^{-\beta}} + \frac{e^{-\beta}}{1+e^{-\beta}}\right) = \frac{2}{\lambda}. \tag{3.5.16}$$

Taking expectation of 3.5.10, plugging in 3.5.11, 3.5.12 and 3.5.16, we get:

$$\mathbb{E}_{LB}\left(\frac{\partial}{\partial \boldsymbol{\theta}}\ell_I(\boldsymbol{\theta})\right) = \mathbf{0}. \tag{3.5.17}$$

Having verified this fact, we move on to the information matrix, which is given by

$$\mathfrak{I}_{I,n}(\boldsymbol{\theta}) = -\mathbb{E}_{LB}\left(\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\ell_I(\boldsymbol{\theta})\right) = \mathbb{E}_{LB}\begin{pmatrix} \sum_{i=1}^{n} z_i^2 e^{\beta z_i}\lambda x_i & \sum_{i=1}^{n} z_i e^{\beta z_i} x_i \\ \sum_{i=1}^{n} z_i e^{\beta z_i} x_i & \sum_{i=1}^{n} \frac{2}{\lambda^2} \end{pmatrix} \tag{3.5.18}$$

Again, this expectation is taken with respect to the sampling distribution of (X, Z), i.e., $f_{LB}(x|z;\beta,\lambda)p_B(z;\lambda)\,dx$. This time, computations are easier, as we see that

$$\begin{aligned} \mathbb{E}_{LB}(Z^2 e^{\beta Z}\lambda X) &= \int_0^\infty \sum_{z=0}^{1} z^2 e^{\beta z}\lambda x f_{LB}(x|z;\beta,\lambda)p_B(z;\beta)\,dx \\ &= \frac{1}{1+e^\beta}\int_0^\infty e^\beta \lambda x f_{LB}(x|1;\beta,\lambda)\,dx \\ &= \frac{2}{1+e^\beta}, \end{aligned} \tag{3.5.19}$$

using the same trick as in 3.5.12, and

$$\begin{aligned} \mathbb{E}_{LB}(Z e^{\beta Z} X) &= \frac{\mathbb{E}_{LB}(Z e^{\beta Z}\lambda X)}{\lambda} \\ &= \frac{2}{(1+e^\beta)\lambda}, \end{aligned} \tag{3.5.20}$$

once more using 3.5.12. In the end,we obtain:

$$\mathfrak{I}_{I,n}(\boldsymbol{\theta}) = n \begin{pmatrix} \frac{2}{1+e^{\beta}} & \frac{2}{(1+e^{\beta})\lambda} \\ \frac{2}{(1+e^{\beta})\lambda} & \frac{2}{\lambda^2} \end{pmatrix}. \tag{3.5.21}$$

The variance of $\hat{\boldsymbol{\theta}}_I$ is given by the inverse of the information matrix:

$$Var(\hat{\boldsymbol{\theta}}_{I,n}) = \mathfrak{I}_{I,n}^{-1}(\boldsymbol{\theta}) = \frac{1}{n} \begin{pmatrix} \frac{(1+e^{\beta})(1+e^{-\beta})}{2} & -\frac{\lambda(1+e^{-\beta})}{2} \\ -\frac{\lambda(1+e^{-\beta})}{2} & \frac{\lambda^2(1+e^{-\beta})}{2} \end{pmatrix}. \tag{3.5.22}$$

One can take the determinant of the matrix above:

$$\det Var(\hat{\boldsymbol{\theta}}_{I,n}) = \frac{\lambda^2(1+e^{\beta})(1+e^{-\beta})}{4n^2}. \tag{3.5.23}$$

The joint likelihood, knowing the true distribution of the covariates, is given by:

$$\begin{aligned} \mathcal{L}_J(\mathbf{x}, \mathbf{z}; \beta, \lambda) &= \prod_{i=1}^{n} \frac{f_U(x_i|z_i; \beta, \lambda)}{\mu(\beta, \lambda)} \\ &= \prod_{i=1}^{n} \frac{e^{\beta z_i} \lambda \exp(-e^{\beta z_i} \lambda x_i)}{\frac{1+e^{-\beta}}{2\lambda}} \\ &= \prod_{i=1}^{n} e^{\beta z_i} \lambda^2 \exp(-e^{\beta z_i} \lambda x_i) \frac{2}{1+e^{-\beta}}. \end{aligned} \tag{3.5.24}$$

The log-likelihood is:

$$\ell_J(\mathbf{x},\mathbf{z};\beta,\lambda)=\sum_{i=1}^{n}\left(\beta z_i+2\log\lambda-e^{\beta z_i}\lambda x_i+\log 2-\log(1+e^{-\beta})\right). \tag{3.5.25}$$

Taking the derivative once to get the score function:

$$\frac{\partial}{\partial\boldsymbol{\theta}}\ell_J(\boldsymbol{\theta})=\begin{pmatrix}\sum_{i=1}^{n}\left(z_i-z_ie^{\beta z_i}\lambda x_i+\frac{1}{1+e^{\beta}}\right)\\ \sum_{i=1}^{n}\left(\frac{2}{\lambda}-e^{\beta z_I}x_i\right)\end{pmatrix}. \tag{3.5.26}$$

Using 3.5.11 and 3.5.12, it is clear that

$$\mathbb{E}_{LB}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\ell_J(\boldsymbol{\theta})\right)=\mathbf{0}. \tag{3.5.27}$$

The information matrix is given by minus the expectation over the sampling distribution of the second derivative of the log-likelihood:

$$\begin{aligned}\mathfrak{I}_{J,n}&=-\mathbb{E}_{LB}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\ell_J(\boldsymbol{\theta})\right)\\&=\mathbb{E}_{LB}\begin{pmatrix}\sum_{i=1}^{n}\left(z_i^2e^{\beta z_i}\lambda x_i+\frac{1}{(1+e^{\beta})(1+e^{-\beta})}\right)&\sum_{i=1}^{n}z_ie^{\beta z_i}x_i\\ \sum_{i=1}^{n}z_ie^{\beta z_i}x_i&\sum_{i=1}^{n}\frac{2}{\lambda^2}\end{pmatrix}\\&=n\begin{pmatrix}\frac{2}{1+e^{\beta}}+\frac{1}{(1+e^{\beta})(1+e^{-\beta})}&\frac{2}{(1+e^{\beta})\lambda}\\ \frac{2}{(1+e^{\beta})\lambda}&\frac{2}{\lambda^2}\end{pmatrix}\end{aligned} \tag{3.5.28}$$

Note that this is very similar to 3.5.21 but with an extra, positive term in the first diagonal element of the matrix. Notice that this "gain" in information is maximized at $\beta = 0$. To get the variance of the MLE, we invert $\mathfrak{I}_{J.n}(\boldsymbol{\theta})$:

$$Var(\hat{\boldsymbol{\theta}}_{J,n}) = \mathfrak{I}_{J,n}^{-1}(\boldsymbol{\theta}) = \frac{1}{n}\begin{pmatrix} \frac{(1+e^{\beta})(1+e^{-\beta})}{3} & -\frac{\lambda(1+e^{-\beta})}{3} \\ -\frac{\lambda(1+e^{-\beta})}{3} & \frac{\lambda^2(3+2e^{-\beta})}{6} \end{pmatrix}. \tag{3.5.29}$$

The determinant of 3.5.29 is given by

$$\det Var(\hat{\boldsymbol{\theta}}_{J,n}) = \frac{\lambda^2(1+e^{\beta})(1+e^{-\beta})}{6n^2}. \tag{3.5.30}$$

which is clearly smaller than 3.5.23. In fact, the relative efficiency for the β, covariate effect is

$$\frac{Var(\hat{\beta}_{I,n})}{Var(\hat{\beta}_{J,n})} = 1.5, \tag{3.5.31}$$

while the relative efficiency for the rate parameter is a function of β:

$$\frac{Var(\hat{\lambda}_{I,n})}{Var(\hat{\lambda}_{J,n})} = \frac{3(1+e^{-\beta})}{3+2e^{-\beta}}, \tag{3.5.32}$$

which is bounded below by 1 as $\beta \longrightarrow \infty$ and bounded above by 1.5 as $\beta \longrightarrow -\infty$. Similarly the ratio of 3.5.23 over 3.5.30 is 1.5.

Chapter 4

Asymptotic Properties of the MLE from $\mathcal{L}_J(\boldsymbol{\theta})$

Now that the proper joint likelihood has been established, it is imperative to derive the asymptotic properties of the resulting estimator.

4.1 Consistency of the MLE

In this section, we establish the consistency of the MLE of $\boldsymbol{\theta}$. It is shown that the MLE is consistent *even if all the observations are censored.* This property holds whether covariates are included or not, as this is possible because the censoring mechanism is informative, and censored observations contain "extra" information about the unknown parameters. It is well known, however, that the MLE based only on censored observations is not consistent under non-informative censoring.

Let the observed data be $y_1, y_2, \cdots, y_n$ where $y_i = (T_i, R_i \wedge C_i, \mathbf{Z}_i, \delta_i)$. We shall use $\mathcal{L}$, dropping the subscript J, to denote the likelihood for this section and the next. Then using 3.2.19,

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \Big(\frac{f_U(u_i, \mathbf{z}_i; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\Big)^{\delta_i} \Big(\int_{\omega \geq v_i} \frac{f_U(\omega, \mathbf{z}_i; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} d\omega\Big)^{1-\delta_i}$$

where $u_i = t_i + r_i$ and $v_i = t_i + c_i$. Let f_{LB} be the density function of F_{LB} and define

$$h(v, \mathbf{z}; \boldsymbol{\theta}) = \int_{\omega \geq v} \frac{f_U(\omega, \mathbf{z}; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} d\omega = \int_{\omega \geq v} \frac{1}{\omega} F_{LB}(d\omega, d\mathbf{z}; \boldsymbol{\theta}) \ .$$

The density function h is called the forward recurrence time density. It is now evident that

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &\propto \prod_{i=1}^{n} F_{LB}^{\delta_i}(du_i, d\mathbf{z}_i; \boldsymbol{\theta}) \Big[\int_{\omega \geq v_i} \frac{1}{\omega} F_{LB}(d\omega, d\mathbf{z}; \boldsymbol{\theta})\Big]^{1-\delta_i} && (4.1.1) \\ &= \prod_{i=1}^{n} f_{LB}^{\delta_i}(u_i, \mathbf{z}_i; \boldsymbol{\theta}) h^{1-\delta_i}(v_i, \mathbf{z}_i, \boldsymbol{\theta}) && (4.1.2) \end{aligned}$$

To establish the consistency of the MLEs we adapt the approach used by Ibragimov and Ha'sminskii (1981, pages 35-38). We make the following assumptions:

A.1 The parameter space, $\boldsymbol{\Theta}$ is a bounded open subset of $\mathbb{R}^k$;

A.2 The probability density function, $f_{LB}(x, \mathbf{z}; \boldsymbol{\theta})$ is a continuous function of θ on $\overline{\Theta}$, the closure of Θ, for almost all $x \in \mathcal{X}$, the sample space;

A.3 The density $f_{LB}(x, \mathbf{z}; \boldsymbol{\theta})$ is identifiable with respect to $\boldsymbol{\theta}$;

A.4 For any $\boldsymbol{\theta} \in \overline{\Theta}$, there exists $\varepsilon > 0$ such that

$$\int_{\mathcal{X}} \sup_{\|\zeta\|<\varepsilon} f_{LB}(x, \mathbf{z}; \boldsymbol{\theta} + \zeta) dx d\mathbf{z} < \infty.$$

It should be noted that if A.1 -A.4 hold for f_{LB}, they hold for f_U.

Theorem 2. *Suppose that A.1-A.4 are fulfilled. Let $\hat{\boldsymbol{\theta}}_n$ be the MLE of $\boldsymbol{\theta}$. Then $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}$ almost surely as $n \to \infty$.*

The main idea of the proof is as follows. Define the likelihood ratio,

$$Z_n(\xi) = \prod_{i=1}^{n} \frac{\mathcal{L}_i(\boldsymbol{\theta} + \xi)}{\mathcal{L}_i(\boldsymbol{\theta})}$$

where ξ is located in a small sphere in $\mathcal{U} = \Theta - \boldsymbol{\theta}$, assuming that $\boldsymbol{\theta}$ is the "true" parameter value. Show that with probability approaching zero $\sup Z_n(\xi) \geq 1$ where the supremum is taken over ξ in this sphere. Using a compactness argument show that $\sup_{\xi \in \Theta - \boldsymbol{\theta}} Z_n(\xi) \geq 1$ with probability approaching zero as $n \to \infty$. Finally, using an argument similar to that of Ibragimov and Ha'sminskii(1981, page 36) conclude that $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}$ w.p. 1 as $n \to \infty$.

In order to carry out the above steps two lemmas are needed to establish the key bound,

$$\mathbb{E}_{\boldsymbol{\theta}}[\sup_{\Gamma} Z_n^{1/2}(\xi)] \leq \exp\Big\{ -n\Big(\frac{1}{2}[k_{\boldsymbol{\theta}}^{f_{LB}}(\frac{\gamma}{2}) + k_{\boldsymbol{\theta}}^{h}(\frac{\gamma}{2})] - \omega_{\boldsymbol{\theta}+\xi_0}^{h}(\varepsilon) - \omega_{\boldsymbol{\theta}+\xi_0}^{f_{LB}}(\varepsilon)\Big)\Big\}.$$

for positive k's and ω's that tend to zero. This bound, in turn, opens the way for an application of Markov's inequality to complete the proof along the lines of Ibragimov and Ha'sminskii.

Proof of Theorem 2. We first established the two lemmas mentioned.

Lemma 1. *Suppose assumptions A1-A4 hold, then*

a) For all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and all $\gamma > 0$,

$$\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|>\gamma} r_{2,f_{LB},\phi}^2(\boldsymbol{\theta},\boldsymbol{\theta}') = \inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|>\gamma} \int_{\mathcal{X}} (f_{LB}^{1/2}(x;\boldsymbol{\theta}) - f_{LB}^{1/2}(x;\boldsymbol{\theta}'))^2 \phi(x) dx d\mathbf{z} = k_{\boldsymbol{\theta}}^{f_{LB}}(\gamma) > 0$$

where $\phi(x) = \frac{1}{x}\int_0^x S_C(\omega)d\omega$, and $S_C(\omega)$ is the survival function of the censoring random variable C. Note that $\phi(x) = P(\delta = 1|T + R = x)$.

b) For all $\boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}$,

$$\Big(\int_{\mathcal{X}} \sup_{\|\zeta\|\leq\varepsilon} \big[f_{LB}^{1/2}(x,\mathbf{z};\boldsymbol{\theta}+\zeta) - f_{LB}^{1/2}(x,\mathbf{z};\boldsymbol{\theta})\big]^2 \phi(x) dx d\mathbf{z}\Big)^{1/2} = \omega_{\boldsymbol{\theta}}^{f_{LB}}(\varepsilon) \to 0 \quad \text{as } \varepsilon \to 0.$$

Proof of Lemma 1. We first notice that the Dominated Convergence Theorem (DCT), A.2 and A.4 imply that $r_{2,f_{LB},\phi}^2(\boldsymbol{\theta},\boldsymbol{\theta}+\zeta)$ is a continuous function of ζ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Part (a) then follows from A.3 and the compactness of

$\overline{\Theta}$. On the other hand, A.2 implies that for any $\boldsymbol{\theta} \in \Theta$,

$$\sup_{\|\zeta\| \leq \varepsilon} \left[f_{LB}^{1/2}(x, \mathbf{z}; \boldsymbol{\theta} + \zeta) - f_{LB}^{1/2}(x, \mathbf{z}; \boldsymbol{\theta}) \right]^2 \longrightarrow 0 \quad \text{as } \varepsilon \longrightarrow 0$$

for all $(x, \mathbf{z}) \in \mathcal{X}$. A.4 and the DCT now complete the proof of (b). □

Lemma 2. *Suppose A.1-A.4 hold for f_{LB}. Then the same conditions holds for h too.*

Proof of Lemma 2. A.1 is automatically fulfilled. To verify A.2 we notice that f_{LB} is a density function and

$$\frac{f_{LB}(s, \mathbf{z}; \boldsymbol{\theta})}{s} I_{[y,\infty)}(s) \leq \frac{f_{LB}(s, \mathbf{z}; \boldsymbol{\theta})}{y}, \qquad \forall s \geq 0.$$

The DCT therefore implies the continuity of $h(x, \mathbf{z}; \boldsymbol{\theta})$ for almost all $x \in \mathcal{X}$. By A.3, $g(x, \mathbf{z}; \boldsymbol{\theta})$ is identifiable with respect to $\boldsymbol{\theta}$. To verify identifiability of h, suppose

$$h(y, \mathbf{z}; \boldsymbol{\theta}) = h(y, \mathbf{z}; \boldsymbol{\theta}') \qquad \forall (y, \mathbf{z}),$$

Then $\frac{\partial}{\partial y} h(y, \mathbf{z}; \boldsymbol{\theta}) = \frac{\partial}{\partial y} h(y; \boldsymbol{\theta}')$ which results in

$$\frac{f_{LB}(y, \mathbf{z}; \boldsymbol{\theta})}{y} = \frac{f_{LB}(y, \mathbf{z}; \boldsymbol{\theta}')}{y} \qquad \forall (y, \mathbf{z}) \text{ s.t. } y \neq 0$$

and hence $f_{LB}(y, \mathbf{z}; \boldsymbol{\theta}) = f_{LB}(y, \mathbf{z}; \boldsymbol{\theta}')$, $\forall (y, \mathbf{z})$, which implies $\boldsymbol{\theta} = \boldsymbol{\theta}'$.

To prove A.4 holds for h, observe that

$$\sup_{\|\zeta\|<\varepsilon} h(y, \mathbf{z}; \boldsymbol{\theta}+\zeta) \leq \int_{s \geq y} \sup_{\|\zeta\|<\varepsilon} \frac{f_{LB}(s, \mathbf{z}; \boldsymbol{\theta}+\zeta)}{s} ds.$$

Thus,

$$\begin{aligned}
\int_{\mathcal{X}} \sup_{\|\zeta\|<\varepsilon} h(y, \mathbf{z}; \boldsymbol{\theta}+\zeta) dy d\mathbf{z} &\leq \int_{\mathcal{X}} \int_{s \geq y} \sup_{\|\zeta\|<\varepsilon} \frac{f_{LB}(s, \mathbf{z}; \boldsymbol{\theta}+\zeta)}{s} ds dy d\mathbf{z} \\
&= \int_0^{\infty} \int_0^{s} \sup_{\|\zeta\|<\varepsilon} \frac{g(s, \mathbf{z}; \boldsymbol{\theta}+\zeta)}{s} dy d\mathbf{z} ds \\
&= \int_0^{\infty} \sup_{\|\zeta\|<\varepsilon} \frac{f_{LB}(s, \mathbf{z}; \boldsymbol{\theta}+\zeta)}{s} \int_0^{s} dy d\mathbf{z} ds \\
&= \int_0^{\infty} \int_{\mathbf{z}} \sup_{\|\zeta\|<\varepsilon} f_{LB}(s, \mathbf{z}; \boldsymbol{\theta}+\zeta) d\mathbf{z} ds < \infty \quad \text{(using A.4)}
\end{aligned}$$

Now the DCT completes the verification of A.4 for h. $\square$

Corollary 1. *If Assumptions A1-A4 hold, then*

a) For all $\boldsymbol{\theta} \in \Theta$ and all $\gamma > 0$,

$$\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|>\gamma} r^2_{2,h,F_C}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|>\gamma} \int_{\mathcal{X}} (h^{1/2}(x, \mathbf{z}; \boldsymbol{\theta}) - h^{1/2}(x, \mathbf{z}; \boldsymbol{\theta}'))^2 F_C(x) dx d\mathbf{z} = k^h_{\boldsymbol{\theta}}(\gamma) > 0,$$

where $F_C(x)$ is the distribution function of the random variable T.

b) For all $\boldsymbol{\theta} \in \overline{\Theta}$,

$$\left(\int_{\mathcal{X}} \sup_{\|\zeta\| \leq \varepsilon} [h^{1/2}(x, \mathbf{z}; \boldsymbol{\theta}+\zeta) - h^{1/2}(x, \mathbf{z}; \boldsymbol{\theta})]^2 F_C(x) dx d\mathbf{z} \right)^{1/2} = \omega^h_{\boldsymbol{\theta}}(\varepsilon) \to 0 \quad \text{as } \varepsilon \to 0.$$

Proof. The result follows immediately from Lemmas 1 and 2.□ □

Details of the proof of Theorem 2.

We can now establish consistency of the MLE. Let Γ be a sphere of small radius ε situated in its entirety in the region $|\xi| > \frac{1}{2}\gamma$. Suppose ξ_0 is the center of Γ. Define

$$Z_n(\xi) = \prod_{i=1}^{n} \frac{\mathcal{L}_i(\boldsymbol{\theta}+\xi)}{\mathcal{L}_i(\boldsymbol{\theta})} \tag{4.1.3}$$

where

$$\mathcal{L}_i(\boldsymbol{\theta}) = \Big(\frac{f_U(u_i, \mathbf{z}_i; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\Big)^{\delta_i}\Big(\int_{\omega \geq v_i} \frac{f_U(\omega, \mathbf{z}_i; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} d\omega\Big)^{1-\delta_i}. \tag{4.1.4}$$

Thus

$$\begin{aligned}
\sup_{\xi\in\Gamma} Z_n^{1/2}(\xi) &= \sup_{\xi\in\Gamma} \prod_{i=1}^{n} \Big(\frac{\mathcal{L}_i(\boldsymbol{\theta}+\xi)}{\mathcal{L}_i(\boldsymbol{\theta})}\Big)^{1/2} \\
&\leq \prod_{i=1}^{n} \Big[\sup_{\xi\in\Gamma}\Big(\frac{\mathcal{L}_i(\boldsymbol{\theta}+\xi)}{\mathcal{L}_i(\boldsymbol{\theta})}\Big)^{1/2}\Big] \\
&= \prod_{i=1}^{n} \mathcal{L}_i^{-1/2}(\boldsymbol{\theta}) \prod_{i=1}^{N} \Big[\mathcal{L}_i^{1/2}(\boldsymbol{\theta}+\xi_0) + \sup_{\|\zeta\|\leq\varepsilon} \Big|\mathcal{L}_i^{1/2}(\boldsymbol{\theta}+\xi_0+\zeta) - \mathcal{L}_i^{1/2}(\theta+\xi_0)\Big|\Big].
\end{aligned}$$

Let $\mathcal{L}$ be the generic $\mathcal{L}_i$. It then follows from the above inequality that

$$\mathbb{E}_{\boldsymbol{\theta}}\Big[\sup_{\xi\in\Gamma} Z_n^{1/2}(\xi)\Big] \leq \Big(\mathbb{E}_{\boldsymbol{\theta}}\Big\{\mathcal{L}^{-1/2}(\boldsymbol{\theta})\Big[\mathcal{L}^{1/2}(\boldsymbol{\theta}+\xi_0) + \sup_{\|\zeta\|\leq\varepsilon}\Big|\mathcal{L}^{1/2}(\boldsymbol{\theta}+\xi_0+\zeta) - \mathcal{L}^{1/2}(\boldsymbol{\theta}+\xi_0)\Big|\Big]\Big\}\Big)^n.$$

On the other hand,

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}}\Big\{\mathcal{L}^{-1/2}(\boldsymbol{\theta})\Big[\mathcal{L}^{1/2}(\boldsymbol{\theta}+\xi_0)+\sup_{\|\zeta\|\le\varepsilon}\Big|\mathcal{L}^{1/2}(\boldsymbol{\theta}+\xi_0+\zeta)-\mathcal{L}^{1/2}(\boldsymbol{\theta}+\xi_0)\Big|\Big]\Big\} \\
&= \mathbb{E}_{\boldsymbol{\theta}}\Big[\mathcal{L}^{-1/2}(\boldsymbol{\theta})\mathcal{L}^{1/2}(\boldsymbol{\theta}+\xi_0)\Big]+\mathbb{E}_{\boldsymbol{\theta}}\Big[\mathcal{L}^{-1/2}(\boldsymbol{\theta})\sup_{\|\zeta\|\le\varepsilon}\Big|\mathcal{L}^{1/2}(\boldsymbol{\theta}+\xi_0+\zeta)-\mathcal{L}^{1/2}(\boldsymbol{\theta}+\xi_0)\Big|\Big] \\
&= I+II\ , \quad \text{say.}
\end{aligned}
$$

We simplify I & II and obtain bounds for them in terms of $k^l_{\boldsymbol{\theta}}(\gamma)$ and $\omega^l_{\boldsymbol{\theta}}(\delta)$ for $l = f_{LB}, h$.

$$
I=\mathbb{E}_{\boldsymbol{\theta}}[\mathcal{L}^{-1/2}(\boldsymbol{\theta})\mathcal{L}^{1/2}(\boldsymbol{\theta}+\xi_0)]=\int_{\mathcal{X}}\mathcal{L}^{-1/2}(\theta)\mathcal{L}^{1/2}(\boldsymbol{\theta}+\xi_0)\lambda_{\boldsymbol{\theta}}(dx) \qquad (4.1.5)
$$

where $\mathcal{X}$ is the sample space defined by the quadruple $X=(T,R,C,\mathbf{Z})$ and $\lambda(dx)$ is the density of X. Note that the definition of the sample points as $X=(T,R,C,\mathbf{Z})$ rather than as $X=(T+R,C,\mathbf{Z})$ is precisely what facilitates our proof.

First,

$$
\begin{aligned}
&\int_{\mathcal{X}}\mathcal{L}^{-1/2}(\boldsymbol{\theta})\mathcal{L}^{1/2}(\boldsymbol{\theta}+\xi_0)\lambda(dx) \\
&= \int_{\mathbf{z}}\int_t\int_r\int_c\left(\frac{f_U(t+r,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\right)^{-\delta/2}\left(\int_{\omega\ge t+c}\frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}d\omega\right)^{-(1-\delta)/2} \\
&\left(\frac{f_U(t+r,\mathbf{z};\boldsymbol{\theta}+\xi_0)}{\mu(\boldsymbol{\theta}+\xi_0)}\right)^{\delta/2}\left(\int_{\omega\ge t+c}\frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta}+\xi_0)}{\mu(\boldsymbol{\theta}+\xi_0)}d\omega\right)^{(1-\delta)/2}\lambda_{\boldsymbol{\theta}}(t,r,c,\mathbf{z})dc\,dr\,dt\,d\mathbf{z}, \\
&= \int_{\mathbf{z}}\int_t\int_r\int_c M(t,r,c,\mathbf{z};\boldsymbol{\theta},\xi_0)dc\,dr\,dt\,d\mathbf{z}\ , \quad \text{say}\ ,
\end{aligned}
$$

where δ is a generic δ_i, and

$$\lambda_{\boldsymbol{\theta}}(t,r,c,\mathbf{z}) = f_{T,R,\mathbf{Z}}(t,r,\mathbf{z};\boldsymbol{\theta})f_C(c) = \frac{f_U(t+r,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}f_C(c)\cdot$$

We then have

$$\begin{aligned}
&\int_{\mathcal{X}} \mathcal{L}^{-1/2}(\boldsymbol{\theta})\mathcal{L}^{1/2}(\boldsymbol{\theta}+\xi_0)\lambda(dx) \\
&= \int_{\mathbf{z}}\int_t\int_r\int_{c\le r} M(t,r,c,\mathbf{z};\boldsymbol{\theta},\xi_0)dcdrdtd\mathbf{z} \\
&\quad + \int_{\mathbf{z}}\int_t\int_r\int_{c>r} M(t,r,c,\mathbf{z};\boldsymbol{\theta},\xi_0)dcdrdtd\mathbf{z} \\
&= \int_{\mathbf{z}}\int_t\int_r\int_{c\le r}\left(\int_{\omega\ge t+c}\frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}d\omega\right)^{-1/2}\left(\int_{\omega\ge a+t}\frac{f(\omega;\theta+\xi_0)}{\mu(\theta+\xi_0)}d\omega\right)^{1/2} \\
&\quad \frac{f_U(t+r,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}f_C(c)dcdrdtd\mathbf{z} \\
&\quad + \int_{\mathbf{z}}\int_t\int_r\int_{c>r}\left(\frac{f_U(t+r,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\right)^{-1/2}\left(\frac{f_U(t+r,\mathbf{z};\boldsymbol{\theta}+\xi_0)}{\mu(\boldsymbol{\theta}+\xi_0)}\right)^{1/2} \\
&\quad \frac{f_U(t+r,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}f_C(c)dcdrdtd\mathbf{z} \\
&= A+B\ , \quad \text{say.}
\end{aligned}$$

Now

$$\begin{aligned}
A &= \int_{\mathbf{z}}\int_t\int_c\Big(\int_{\omega\ge t+c}\frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}d\omega\Big)^{-1/2}\Big(\int_{\omega\ge t+c}\frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta}+\xi_0)}{\mu(\boldsymbol{\theta}+\xi_0)}d\omega\Big)^{1/2}\times \\
&\quad \int_{r\ge c}\frac{f_U(t+r,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}dr\,f_C(c)dcdtd\mathbf{z} \\
&= \int_{\mathbf{z}}\int_t\int_c\Big(\int_{\omega\ge t+c}\frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}d\omega\Big)^{1/2}\Big(\int_{\omega\ge t+c}\frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta}+\xi_0)}{\mu(\boldsymbol{\theta}+\xi_0)}d\omega\Big)^{1/2}f_C(c)dcdtd\mathbf{z}.
\end{aligned}$$

Define $v = t + c$ and $w = c$. Then

$$\begin{aligned} A &= \int_{\mathbf{z}}\int_{v}\int_{w\le v}\Big(\int_{\omega\ge v}\frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}d\omega\Big)^{1/2}\Big(\int_{\omega\ge v}\frac{f_U(\omega;\boldsymbol{\theta}+\xi_0)}{\mu(\boldsymbol{\theta}+\xi_0)}d\omega\Big)^{1/2} f_C(w)dwdvd\mathbf{z} \\ &= \int_{\mathbf{z}}\int_{v} h^{1/2}(v,\mathbf{z};\boldsymbol{\theta})h^{1/2}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0)F_C(v)dvd\mathbf{z}\ \cdot \end{aligned}$$

Since

$$B = \int_{\mathbf{z}}\int_{t}\int_{r}\Big(\frac{f_U(t+r,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\Big)^{1/2}\Big(\frac{f_U(t+r,\mathbf{z};\boldsymbol{\theta}+\xi_0)}{\mu(\boldsymbol{\theta}+\xi_0)}\Big)^{1/2} S_C(r)drdtd\mathbf{z}\ ,$$

we may let $v = t + r$ and $w = r$ to obtain,

$$\begin{aligned} B &= \int_{\mathbf{z}}\int_{v}\int_{w\le v}\Big(\frac{f_U(v,\mathbf{z};\boldsymbol{\theta})}{\mu(\theta)}\Big)^{1/2}\Big(\frac{f(v,\mathbf{z};\boldsymbol{\theta}+\xi_0)}{\mu(\boldsymbol{\theta}+\xi_0)}\Big)^{1/2} S_C(w)dwdvd\mathbf{z} \\ &= \int_{\mathbf{z}}\int_{v}\Big(\frac{f_U(v,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\Big)^{1/2}\Big(\frac{f_U(v,\mathbf{z};\boldsymbol{\theta}+\xi_0)}{\mu(\boldsymbol{\theta}+\xi_0)}\Big)^{1/2}\int_0^v S_C(w)dwdvd\mathbf{z} \\ &= \int_{\mathbf{z}}\int_{v} f_{LB}^{1/2}(v,\mathbf{z};\boldsymbol{\theta})f_{LB}^{1/2}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0)\frac{1}{v}\int_0^v S_C(w)dwdvd\mathbf{z}, \end{aligned}$$

where $f_{LB}(v,\mathbf{z};\boldsymbol{\theta}) = \frac{vf_U(v,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}$ is the length-biased density. Thus

$$\begin{aligned} I &= A + B \\ &= \int_{\mathbf{z}}\int_{v} h^{1/2}(v,\mathbf{z};\boldsymbol{\theta})h^{1/2}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0)F_C(v)dvd\mathbf{z} + \\ &\quad \int_{\mathbf{z}}\int_{v} f_{LB}^{1/2}(v,\mathbf{z};\boldsymbol{\theta})f_{LB}^{1/2}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0)\frac{1}{v}\int_0^v S_C(\omega)d\omega dvd\mathbf{z}. \end{aligned}$$

On the other hand,

$$\begin{aligned} A &= \frac{1}{2}\Big(\int_{\mathbf{z}}\int_{v} h(v,\mathbf{z};\boldsymbol{\theta})F_C(v)dvd\mathbf{z} + \int_{\mathbf{z}}\int_{v} h(v,\mathbf{z};\boldsymbol{\theta}+\xi_0)F_C(v)dvd\mathbf{z} \\ &\quad - \int_{\mathbf{z}}\int_{v}\Big(h^{1/2}(v,\mathbf{z};\boldsymbol{\theta}) - h^{1/2}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0)\Big)^2 F_C(v)dvd\mathbf{z}\Big) \end{aligned}$$

and

$$\begin{aligned} B &= \int_{\mathbf{z}}\int_{v} f_{LB}^{1/2}(v,\mathbf{z};\boldsymbol{\theta}) f_{LB}^{1/2}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0)\frac{1}{v}\int_0^v S_C(\omega)d\omega dvd\mathbf{z} \\ &= \frac{1}{2}\Big(\int_{\mathbf{z}}\int_{v} f_{LB}(v,\mathbf{z};\boldsymbol{\theta})\frac{1}{v}\int_0^v S_C(\omega)d\omega dvd\mathbf{z} \\ &\quad + \int_{\mathbf{z}}\int_{v} f_{LB}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0)\frac{1}{v}\int_0^v S_C(\omega)d\omega dvd\mathbf{z} \\ &\quad - \int_{\mathbf{z}}\int_{v}\Big(f_{LB}^{1/2}(v,\mathbf{z};\boldsymbol{\theta}) - f_{LB}^{1/2}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0)\Big)^2\frac{1}{v}\int_0^v S_C(\omega)d\omega dvd\mathbf{z}\Big). \end{aligned}$$

We now note that

$$\begin{aligned} &\int_{\mathbf{z}}\int_{v} h(v,\mathbf{z};\boldsymbol{\theta})F_C(v)dvd\mathbf{z} + \int_{\mathbf{z}}\int_{v} f_{LB}(v,\mathbf{z};\boldsymbol{\theta})\frac{1}{v}\int_0^v S_C(\omega)d\omega dvd\mathbf{z} \\ &= \int_{\mathbf{z}}\int_{v}\int_{\omega\geq v}\frac{1}{\omega}f_{LB}(\omega,\mathbf{z};\boldsymbol{\theta})F_C(v)d\omega dvd\mathbf{z} + \int_{\mathbf{z}}\int_{v}\int_{\omega<v}\frac{1}{v}f_{LB}(v,\mathbf{z};\boldsymbol{\theta})S_C(\omega)d\omega dvd\mathbf{z} \\ &= \int_{\mathbf{z}}\int_{v}\int_{\omega\geq v}\frac{1}{\omega}f_{LB}(\omega,\mathbf{z};\boldsymbol{\theta})F_C(v)d\omega dvd\mathbf{z} + \int_{\mathbf{z}}\int_{\omega}\int_{v\geq\omega}\frac{1}{v}f_{LB}(v,\mathbf{z};\boldsymbol{\theta})S_C(\omega)dvd\omega d\mathbf{z} \\ &= \int_{\mathbf{z}}\int_{\omega}\int_{v\geq\omega}\frac{1}{v}f_{LB}(v,\mathbf{z};\boldsymbol{\theta})dvd\omega d\mathbf{z} \\ &= \int_{\mathbf{z}}\int_{\omega} h(\omega,\mathbf{z};\theta)d\omega d\mathbf{z} \\ &= 1. \end{aligned}$$

This then implies that

$$
\begin{aligned}
I &= A+B \\
&= \frac{1}{2}\Big(2-\int_{\mathbf{z}}\int_{v}\Big(h^{1/2}(v,\mathbf{z};\boldsymbol{\theta})-h^{1/2}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0)\Big)^2 F_C(v)dvd\mathbf{z} \\
&\quad -\int_{\mathbf{z}}\int_{v}\Big(f_{LB}^{1/2}(v,\mathbf{z};\boldsymbol{\theta})-f_{LB}^{1/2}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0)\Big)^2\frac{1}{v}\int_0^v S_C(\omega)d\omega dvd\mathbf{z}\Big) \\
&= 1-\frac{k_{\boldsymbol{\theta}}^{f_{LB}}(\frac{\gamma}{2})+k_{\boldsymbol{\theta}}^{h}(\frac{\gamma}{2})}{2}, \quad \text{(using Lemma 1 and Lemma 2).} \qquad (4.1.6)
\end{aligned}
$$

In similar fashion it may be shown that

$$
\begin{aligned}
II &= \mathbb{E}_{\boldsymbol{\theta}}\Big[\mathcal{L}^{-1/2}(\boldsymbol{\theta})\sup_{\|\zeta\|\le\varepsilon}\Big|\mathcal{L}^{1/2}(\boldsymbol{\theta}+\xi_0+\zeta)-\mathcal{L}^{1/2}(\boldsymbol{\theta}+\xi_0)\Big|\Big] \\
&= \int_{\mathcal{X}}\mathcal{L}^{-1/2}(\boldsymbol{\theta})\sup_{\|\zeta\|\le\varepsilon}\Big|\mathcal{L}^{1/2}(\boldsymbol{\theta}+\xi_0+\zeta)-\mathcal{L}^{1/2}(\boldsymbol{\theta}+\xi_0)\Big|\lambda_\theta(dx) \\
&= \int_{\mathbf{z}}\int_t\int_r\int_{c\le r}\underbrace{\Big(\int_{\omega\ge t+c}\frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}d\omega\Big)^{-1/2}}_{h(t+c,\mathbf{z};\boldsymbol{\theta})}\sup_{\|\zeta\|\le\varepsilon}\Big|h^{1/2}(t+c,\mathbf{z};\boldsymbol{\theta}+\xi_0+\zeta) \\
&\quad -h^{1/2}(t+c;\boldsymbol{\theta}+\xi_0)\Big|\frac{f_U(t+c;\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}f_C(c)dcdrdtd\mathbf{z} \\
&\quad +\int_{\mathbf{z}}\int_t\int_r\int_{c>r}\Big(\frac{f_U(t+c,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\Big)^{-1/2}\sup_{\|\zeta\|\le\varepsilon}\Big|\Big(\frac{f_U(t+c,\mathbf{z};\boldsymbol{\theta}+\xi_0+\zeta)}{\mu(\boldsymbol{\theta}+\xi_0+\zeta)}\Big)^{1/2} \\
&\quad -\Big(\frac{f_U(t+c,\mathbf{z};\boldsymbol{\theta}+\xi_0)}{\mu(\boldsymbol{\theta}+\xi_0)}\Big)^{1/2}\Big|\frac{f_U(t+c,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}f_C(c)dcdrdtd\mathbf{z} \\
&= C+D, \quad \text{say.}
\end{aligned}
$$

We now simplify C and D as follows.

$$
\begin{aligned}
C &= \int_{\mathbf{z}}\int_t\int_r\int_{c\le r}\Big(\int_{\omega\ge t+c}\frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}d\omega\Big)^{-1/2}\sup_{\|\zeta\|\le\varepsilon}\Big|h^{1/2}(t+c,\mathbf{z};\boldsymbol{\theta}+\xi_0+\zeta)\\
&\qquad -h^{1/2}(t+c,\mathbf{z};\boldsymbol{\theta}+\xi_0)\Big|\frac{f_U(t+r,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}f_C(c)dcdrdtd\mathbf{z}\\
&= \int_{\mathbf{z}}\int_t\int_c h^{1/2}(t+c,\mathbf{z};\boldsymbol{\theta})\sup_{\|\zeta\|\le\varepsilon}\Big|h^{1/2}(t+c,\mathbf{z};\boldsymbol{\theta}+\xi_0+\zeta)\\
&\qquad -h^{1/2}(t+c,\mathbf{z};\boldsymbol{\theta}+\xi_0)\Big|f_C(c)dcdtd\mathbf{z}\\
&= \int_{\mathbf{z}}\int_v h^{1/2}(v,\mathbf{z};\boldsymbol{\theta})\sup_{\|\zeta\|\le\varepsilon}\Big|h^{1/2}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0+\zeta)-h^{1/2}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0)\Big|F_C(v)dvd\mathbf{z}
\end{aligned}
$$

by setting $v=t+c$ and $w=c$. Next,

$$
\begin{aligned}
D &= \int_{\mathbf{z}}\int_t\int_r f_{LB}^{1/2}(t+r,\mathbf{z};\boldsymbol{\theta})\sup_{\|\zeta\|\le\varepsilon}\Big|f_{LB}^{1/2}(t+r,\mathbf{z};\boldsymbol{\theta}+\xi_0+\zeta)\\
&\qquad -f_{LB}^{1/2}(t+r,\mathbf{z};\boldsymbol{\theta}+\xi_0)\Big|\frac{1}{t+r}S_C(r)drdtd\mathbf{z}\\
&= \int_v f_{LB}^{1/2}(v,\mathbf{z};\boldsymbol{\theta})\sup_{\|\zeta\|\le\varepsilon}\Big|f_{LB}^{1/2}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0+\zeta)\\
&\qquad -f_{LB}^{1/2}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0)\Big|\frac{1}{v}\int_0^v S_C(\omega)d\omega dvd\mathbf{z}.
\end{aligned}
$$

by setting $v=t+r$, and $w=r$.

It then follows from the Cauchy-Schwarz inequality that

$$\begin{aligned}
II &= C + D \\
&= \int_{\mathbf{z}}\int_{v} h^{1/2}(v,\mathbf{z};\boldsymbol{\theta}) \sup_{\|\zeta\|\le\varepsilon} \Big| h^{1/2}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0+\zeta) - h^{1/2}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0) \Big| F_C(v)dvd\mathbf{z} \\
&+ \int_{\mathbf{z}}\int_{v} f_{LB}^{1/2}(v,\mathbf{z};\boldsymbol{\theta}) \sup_{\|\zeta\|\le\varepsilon} \Big| f_{LB}^{1/2}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0+\zeta) - f_{LB}^{1/2}(v,\mathbf{z};\boldsymbol{\theta}+\xi_0) \Big| \frac{1}{v}\int_0^v S_C(t)dtdvd\mathbf{z} \\
&\le \omega^h_{\boldsymbol{\theta}+\xi_0}(\varepsilon) + \omega^{f_{LB}}_{\boldsymbol{\theta}+\xi_0}(\varepsilon) \qquad \text{(using Lemma 1 and Lemma 2)} .
\end{aligned}$$

Using (4.1.6), (4.1.7) and the elementary inequality $1+a \le e^a$, $-\infty \le a \le \infty$, we obtain

$$\mathbb{E}_{\boldsymbol{\theta}}[\sup_{\Gamma} Z_n^{1/2}(\xi)] \le \exp\Big\{ -n\Big(\frac{1}{2}[k^{f_{LB}}_{\boldsymbol{\theta}}(\frac{\gamma}{2}) + k^h_{\boldsymbol{\theta}}(\frac{\gamma}{2})] - \omega_{\boldsymbol{\theta}+\xi_0}(\varepsilon) - \omega_{\boldsymbol{\theta}+\xi_0}(\varepsilon)\Big)\Big\}.$$

The proof is now completed in similar fashion to the proof of Theorem 4.3 of Ibragimov and Has'minskii(1981, page 36-38) and by referring to Remark 4.1 by the same authors. □

4.2 Asymptotic Normality of the MLE

In this section we establish the asymptotic normality of the MLE under appropriate regularity conditions. While the proof relies on the usual approach via a Taylor expansion of the score function, once again, the difficulties posed by informative censoring and covariates need to be resolved.

Define $\psi(x, \mathbf{z}; \boldsymbol{\theta}) = \log f_{LB}(x, \mathbf{z}; \boldsymbol{\theta})$ and let

$$\psi_{pqr}(x, \mathbf{z}; \boldsymbol{\theta}) = \frac{\partial^3}{\partial \boldsymbol{\theta}_p \partial \boldsymbol{\theta}_q \partial \boldsymbol{\theta}_r} \log f_{LB}(x, \mathbf{z}; \boldsymbol{\theta}) \quad .$$

First and second partial derivatives of ψ are accordingly denoted by ψ_p and ψ_{pq}. We make the following assumptions:

A.5

(a) Suppose ψ admits all third order partial derivatives for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

(b) For any $p, q = 1, 2, \ldots, k$, the third moments of $\psi_p(X, \mathbf{Z}; \boldsymbol{\theta})$ and $\psi_{pq}(X, \mathbf{Z}; \boldsymbol{\theta})$ exist for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

(c) There exists $\mathcal{K}(x) = O(e^x)$ as $x \to \infty$, such that

$$\left| \int_{\mathbf{z}} \psi_{pqr}(x, \mathbf{z}; \boldsymbol{\theta}) d\mathbf{z} \right| \leq \mathcal{K}(x) \quad \text{for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

for $p, q, r = 1, 2, \ldots, k$, where

$$E_{\boldsymbol{\theta}}[\mathcal{K}^3(X)] < \infty \quad \text{for all } \boldsymbol{\theta} \in \boldsymbol{\Theta} \quad .$$

A.6

(a) The following equations hold:

$$E_{\boldsymbol{\theta}}[\psi_r(X, \mathbf{Z}; \boldsymbol{\theta})] = 0 \quad \text{for } r = 1, 2, \ldots, k \ ,$$

and

$$\mathfrak{I}_{pq}^{f_{LB}}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\psi_p(X, \mathbf{Z}; \boldsymbol{\theta}) \cdot \psi_q(X, \mathbf{Z}; \boldsymbol{\theta})] = E_{\boldsymbol{\theta}}[-\psi_{pq}(X, \mathbf{Z}; \boldsymbol{\theta})]$$

for $p, q = 1, 2, \ldots, k$.

(b) The information matrix $\mathfrak{I}^l(\boldsymbol{\theta}) = [\mathfrak{I}_{pq}^l(\boldsymbol{\theta})]_{pq=1,2,\cdots,k}$ is positive definite for $l = f_{LB}, h$ and all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Notice that A.6(a) merely concerns the interchanging of "$\int$" and "∂". Assumption A.6(b) is, however, of a different nature. Except for members of an exponential family, checking positive-definiteness of the information matrix can be difficult. It is comforting, however, that identifiability (A.3) along with smoothness, ensures that positive-definiteness of the information matrix rarely fails to hold (Asgharian 2001). On the other hand, using Lemma 2, assumption A.3 holds for h if f_{LB} fulfills the same condition. This then means that A.6(b) holds for h if it holds for f_{LB}. Finally, assumption A.5(b) is needed to establish differentiability of h.

Theorem 3. *Suppose Conditions A.1 and A.3-A.6 are fulfilled. Then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(0, \mathfrak{I}^{-1}) ,$$

where $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$ are, respectively, the true parameter value and its maximum *likelihood estimator. Let U be a generic U_i and $\mathfrak{I} = \mathfrak{I}^1 + \mathfrak{I}^2$, where*

$$\begin{aligned}\mathfrak{I}^1_{lj} &= \operatorname{cov}\Big(\delta \frac{\partial}{\partial \boldsymbol{\theta}_l} \log f_{LB,\boldsymbol{\theta}_0}(U, \mathbf{Z}), \delta \frac{\partial}{\partial \boldsymbol{\theta}_j} \log f_{LB,\boldsymbol{\theta}_0}(U, \mathbf{Z})\Big) \\ &= E\Big[\delta\Big(\frac{\partial}{\partial \boldsymbol{\theta}_l} \log f_{LB,\boldsymbol{\theta}_0}(U, \mathbf{Z})\Big)\Big(\frac{\partial}{\partial \boldsymbol{\theta}_j} \log f_{LB,\boldsymbol{\theta}_0}(U, \mathbf{Z})\Big)\Big]\end{aligned}$$

and

$$\begin{aligned}\mathfrak{I}^2_{lj} &= \operatorname{cov}\Big((1-\delta) \frac{\partial}{\partial \boldsymbol{\theta}_l} \log h_{\boldsymbol{\theta}_0}(U, \mathbf{Z}), (1-\delta) \frac{\partial}{\partial \boldsymbol{\theta}_j} \log h_{\boldsymbol{\theta}_0}(U, \mathbf{Z})\Big) \\ &= E\Big[(1-\delta)\Big(\frac{\partial}{\partial \boldsymbol{\theta}_l} \log h_{\boldsymbol{\theta}_0}(U, \mathbf{Z})\Big)\Big(\frac{\partial}{\partial \boldsymbol{\theta}_j} \log h_{\boldsymbol{\theta}_0}(U, \mathbf{Z})\Big)\Big].\end{aligned}$$

$l, j = 1, ..., k$.

Proof. In the sequel we use the following notations,

$$g_{\boldsymbol{\theta}}(u, \mathbf{z}) = \frac{f_U(u, \mathbf{z}; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}, \qquad h_{\boldsymbol{\theta}}(v, \mathbf{z}) = \int_{\omega \geq v} \frac{f_U(\omega, \mathbf{z}; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} d\omega.$$

Define,

$$\mathcal{U}_i^T(\boldsymbol{\theta}) = \Big(\frac{\partial}{\partial \boldsymbol{\theta}_1} \log g_{\boldsymbol{\theta}}(u_i, \mathbf{z}_i), ..., \frac{\partial}{\partial \boldsymbol{\theta}_k} \log g_{\boldsymbol{\theta}}(u_i, \mathbf{z}_i)\Big),$$

and

$$\mathcal{V}_i^T(\boldsymbol{\theta}) = \Big(\frac{\partial}{\partial \boldsymbol{\theta}_1} \log h_{\boldsymbol{\theta}}(v_i, \mathbf{z}_i), ..., \frac{\partial}{\partial \boldsymbol{\theta}_k} \log h_{\boldsymbol{\theta}}(v_i, \mathbf{z}_i)\Big),$$

for $i = 1, 2, \cdots, n$.

Using the regularity conditions contained in Lemma 3 one may derive the first order Taylor expansions,

$$\mathcal{U}_i^T(\hat{\boldsymbol{\theta}}) = \mathcal{U}_i^T(\boldsymbol{\theta}_0) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{U}_i^T(\boldsymbol{\theta}_0) + O_p\Big(\frac{1}{n}\Big),$$
$$\mathcal{V}_i^T(\hat{\boldsymbol{\theta}}) = \mathcal{V}_i^T(\boldsymbol{\theta}_0) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{V}_i^T(\boldsymbol{\theta}_0) + O_p\Big(\frac{1}{n}\Big),$$

and using the fact that

$$\sum_{i=1}^{n} \Big[\delta_i \mathcal{U}_i^T(\hat{\boldsymbol{\theta}}) + (1-\delta_i)\mathcal{V}_i^T(\hat{\boldsymbol{\theta}})\Big] = 0,$$

we obtain the key equation,

$$\begin{aligned} 0 = {} & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big[\delta_i \mathcal{U}_i^T(\boldsymbol{\theta}_0) + (1-\delta_i)\mathcal{V}_i^T(\boldsymbol{\theta}_0)\Big] \\ & + \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \frac{1}{n} \sum_{i=1}^{n} \Big[\delta_i \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{U}_i^T(\boldsymbol{\theta}_0) + (1-\delta_i)\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{V}_i^T(\boldsymbol{\theta}_0))\Big] + O_p(\frac{1}{\sqrt{n}}). \end{aligned} \tag{4.2.1}$$

Once equation (4.2.1) has been established the main task is to show that

$$E\Big[\frac{\partial}{\partial \boldsymbol{\theta}_j} \log \mathcal{L}_i(\boldsymbol{\theta})\Big] = 0 \ , \tag{4.2.2}$$

find

$$Var\Big[\delta_i \frac{\partial}{\partial \boldsymbol{\theta}_j} \log g_{\boldsymbol{\theta}}(U_i, \mathbf{Z}_i)\Big] \ , \tag{4.2.3}$$

show that

$$\frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{\partial}{\partial\boldsymbol{\theta}}\mathcal{U}_i^T(\boldsymbol{\theta}_0)\xrightarrow{a.s.}Var\Big[\delta\mathcal{U}^T(\boldsymbol{\theta}_0)\Big]\quad\text{as}\quad n\to\infty\ , \tag{4.2.4}$$

and that

$$\frac{1}{n}\sum_{i=1}^{n}\Big[\delta_i\frac{\partial}{\partial\boldsymbol{\theta}}\mathcal{U}_i^T(\boldsymbol{\theta}_0)+(1-\delta_i)\frac{\partial}{\partial\boldsymbol{\theta}}\mathcal{V}_i^T(\boldsymbol{\theta}_0)\Big]\xrightarrow{a.s.}Var\Big[\delta\mathcal{U}^T(\boldsymbol{\theta}_0)+(1-\delta)\mathcal{V}^T(\boldsymbol{\theta}_0)\Big] \tag{4.2.5}$$

as $n\longrightarrow\infty$.

It is in these steps that the informative censoring needs careful consideration. The details are contained in the appendix. The proof is finally completed, using (4.2.1) by applying the Central Limit Theorem to

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\delta_i\mathcal{U}_i^T(\boldsymbol{\theta}_0)+(1-\delta_i)\mathcal{V}_i(\boldsymbol{\theta}_0)\Big]\cdot \tag{4.2.6}$$

$\square$

It is worth noting that

$$\begin{aligned}\mathrm{Var}\Big(\delta\frac{\partial}{\partial\boldsymbol{\theta}_j}\log g_{\boldsymbol{\theta}_0}(U,\mathbf{Z})\Big) &= -E\Big[\delta\Big(\frac{\partial^2}{\partial\boldsymbol{\theta}_j^2}\log g_{\boldsymbol{\theta}_0}(U,\mathbf{Z})\Big)\Big]\\ &= -\int_{\mathbf{z}}\int_r\int_t\frac{\partial^2}{\partial\boldsymbol{\theta}_j^2}\log g_{\boldsymbol{\theta}_0}(t+r,\mathbf{z})S_C(r)g_{\boldsymbol{\theta}_0}(t+r,\mathbf{z})dtdrd\mathbf{z}.\end{aligned}$$

This shows that $\mathrm{Var}(\delta\frac{\partial}{\partial\boldsymbol{\theta}_j}\log g_{\boldsymbol{\theta}_0}(U,\mathbf{Z}))$ depends on the censoring distribution. The dependence is not just through $P(\delta=1)$, i.e., the proportion of censored observations.

Details of the proof of Theorem 3. One lemma is needed in the proof for asymptotic normality.

Lemma 3. *Suppose A.5 and A.6(a) hold for g. Then A.5(a,c) and A.6(a) hold for h.*

Remark. An inspection of the proof of Lemma 3 shows that the condition $\mathcal{K}(x) = O(e^x)$ as $x \to \infty$, imposed on $\mathcal{K}$ in Assumption A.5(c), can be replaced by

$$\mathcal{K}(x) \sim \frac{\int_0^x \mathcal{K}(s)ds}{x} .$$

Proof of Lemma 3. We first note that

$$\frac{\partial^3}{\partial \boldsymbol{\theta}^3}\psi(y, \mathbf{z}; \boldsymbol{\theta}) = \left(\frac{f_{LB}^{(3)}(y, \mathbf{z}; \boldsymbol{\theta})}{f_{LB}(y, \mathbf{z}; \boldsymbol{\theta})}\right) - 3\left(\frac{f_{LB}^{(1)}(y, \mathbf{z}; \boldsymbol{\theta})}{f_{LB}(y, \mathbf{z}; \boldsymbol{\theta})}\right)\left(\frac{f_{LB}^{(2)}(y, \mathbf{z}; \boldsymbol{\theta})}{f_{LB}(y, \mathbf{z}; \boldsymbol{\theta})}\right) + 2\left(\frac{f_{LB}^{(1)}(y, \mathbf{z}; \boldsymbol{\theta})}{f_{LB}(y, \mathbf{z}; \boldsymbol{\theta})}\right)^3 .$$

Using A.5 and a Taylor expansion, the first three derivatives of $f_{LB}(x, \mathbf{z}; \boldsymbol{\theta})$ are bounded by some function of x whose moment exists. Therefore $h(x, \mathbf{z}; \boldsymbol{\theta})$ possesses the first three derivatives and we can interchange $\int$ and ∂. This implies that A.5(a) holds for h.

In order to establish assumption A.5(c) for h we notice that

$$\partial \log h(y, \mathbf{z}; \boldsymbol{\theta})/\partial \boldsymbol{\theta} = \frac{\int_{s \geq y} \frac{1}{s} \frac{\partial}{\partial \boldsymbol{\theta}} f_{LB}(s, \mathbf{z}; \boldsymbol{\theta}) ds}{\int_{s \geq y} \frac{1}{s} f_{LB}(s, \mathbf{z}; \boldsymbol{\theta}) ds}$$

and therefore,

$$\lim_{y \to \infty} \frac{\partial \log h(y, \mathbf{z}; \boldsymbol{\theta})/\partial \boldsymbol{\theta}}{\partial \log f_{LB}(y, \mathbf{z}; \boldsymbol{\theta})/\partial \boldsymbol{\theta}} = 1.$$

This means that $\partial \log h(y;\theta)/\partial\theta \sim \partial \log g(y,\theta)/\partial\theta$ for large values of y. It is then not hard to see that

$$\frac{h^{(k)}(y, \mathbf{z}; \boldsymbol{\theta})}{h(y, \mathbf{z}; \boldsymbol{\theta})} \sim \frac{f_{LB}^{(k)}(y, \mathbf{z}; \boldsymbol{\theta})}{f_{LB}(y, \mathbf{z}; \boldsymbol{\theta})} \qquad \text{as } y \to \infty, \quad \forall k \ .$$

This implies that $\frac{\partial^k}{\partial\theta^k} \log h(y;\theta) \sim \frac{\partial^k}{\partial\theta^k} \log g(y;\theta)$ as $y \to \infty$, $\forall k$. Thus for large values of y, $|\frac{\partial^{(k)}}{\partial\theta^k} \log h(y;\theta)| \leq \mathcal{K}(y)$, if the same condition holds for g, i.e., $|\frac{\partial^{(k)}}{\partial\theta^k} \log g(y;\theta)| \leq \mathcal{K}(y)$.

It remains to show that $E_h[\mathcal{K}(Y)]$ exists if $E_{f_{LB}}[\mathcal{K}(Y)] < \infty$. We have, by the definition of h,

$$\begin{aligned}
E_h[\mathcal{K}(Y)] &= \int_{\mathbf{z}} \int_0^\infty h(y, \mathbf{z}; \boldsymbol{\theta}) \mathcal{K}(y) dy d\mathbf{z} \\
&= \int_{\mathbf{z}} \int_0^\infty \int_{s \geq y} \frac{1}{s} f_{LB}(s, \mathbf{z}; \boldsymbol{\theta}) \mathcal{K}(y) ds dy d\mathbf{z} \\
\text{(Using Fubini)} \quad &= \int_0^\infty \int_{\mathbf{z}} \frac{\int_0^s \mathcal{K}(y) dy}{s} f_{LB}(s, \mathbf{z}; \boldsymbol{\theta}) d\mathbf{z} ds \\
&= \int_0^\infty \int_{\mathbf{z}} \frac{\kappa(s)}{s} f_{LB}(s, \mathbf{z}; \boldsymbol{\theta}) d\mathbf{z} ds \\
&= E_{f_{LB}}[\frac{\kappa(X)}{X}] \ ,
\end{aligned}$$

where

$$\kappa(s) = \int_0^s \mathcal{K}(y) dy \ .$$

Now $\mathcal{K}(y) = O(e^y)$ implies that $\kappa(s)/s = O(e^s)$, and therefore $E_h[\mathcal{K}(Y)] < \infty$. Assumption A.6(a) for h readily follows from Fubini's theorem and the fact that

$$\partial^k h(y, \mathbf{z}; \boldsymbol{\theta})/\partial \boldsymbol{\theta}^k = \int_{s \geq y} \frac{1}{s} \frac{\partial^k}{\partial \boldsymbol{\theta}^k} f_{LB}(s, \mathbf{z}; \boldsymbol{\theta}) ds \quad \text{for} \quad k = 1, 2, 3$$

This completes the proof. □

Completion of the proof of Theorem 3.

The j-th component of the score function is

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}_j} \log \mathcal{L}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}_j} \log \mathcal{L}_i(\boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \Big[\delta_i \frac{\partial}{\partial \boldsymbol{\theta}_j} \log \frac{f_U(u_i, \mathbf{z}_i; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} + (1 - \delta_i) \frac{\partial}{\partial \boldsymbol{\theta}_j} \log \int_{\omega \geq v_i} \frac{f_U(\omega, \mathbf{z}_i; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} d\omega\Big].
\end{aligned}$$

Now taking expectation we obtain

$$\begin{aligned}
\mathbb{E}\Big[\frac{\partial}{\partial \boldsymbol{\theta}_j} \log \mathcal{L}_i(\boldsymbol{\theta})\Big] &= \mathbb{E}\Big[\delta_i \frac{\partial}{\partial \boldsymbol{\theta}_j} \log \frac{f_U(U_i, \mathbf{Z}_i; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} \\
&\quad + (1 - \delta_i) \frac{\partial}{\partial \boldsymbol{\theta}_j} \log \int_{\omega \geq V_i} \frac{f_U(\omega, \mathbf{Z}_i; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} d\omega\Big] \\
&= \mathbb{E}\Big[\frac{\partial}{\partial \boldsymbol{\theta}_j} \log \frac{f_U(U_i, \mathbf{Z}_i; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} | \delta_i = 1\Big] P(\delta = 1) \\
&\quad + \mathbb{E}\Big[\frac{\partial}{\partial \boldsymbol{\theta}_j} \log \int_{\omega \geq V_i} \frac{f_U(\omega, \mathbf{Z}_i; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} d\omega | \delta_i = 0\Big] P(\delta_i = 0).
\end{aligned}$$

We show that $\mathbb{E}\Big[\frac{\partial}{\partial \boldsymbol{\theta}_j} \log \mathcal{L}_i(\boldsymbol{\theta})\Big] = 0$. Define

$$\varsigma = P(\delta = 1) = P(R \le C) = \int_0^\infty F_R(c) f_C(c) dc,$$

Since $C \perp (T, R, \mathbf{Z})$, we have

$$\begin{aligned} dP(x, \mathbf{z}, \delta = 1) &= P\big(T \in (t, t+dt],\ R \wedge C \in (r, r+dr],\ \mathbf{z} \in B_{d\mathbf{z}}(\mathbf{z}),\ \delta = 1\big) \\ &= P\big(T \in (t, t+dt],\ R \in (r, r+dr],\ \mathbf{z} \in B_{d\mathbf{z}}(\mathbf{z}),\ R \le C\big) \\ &= f_{T,R,\mathbf{Z}}(t, r, \mathbf{z}) S_C(r)\ dt dr d\mathbf{z} \\ &= S_C(r) \frac{f_U(t+r, \mathbf{z}; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\ dt dr d\mathbf{z}\ , \end{aligned}$$

which readily implies

$$dP(x, \mathbf{z} \mid \delta = 1) = \frac{S_C(r) f_U(t+r, \mathbf{z}; \boldsymbol{\theta})}{\varsigma\ \mu(\boldsymbol{\theta})}\ dt dr d\mathbf{z}.$$

Then we have (dropping the subscript for convenience):

$$\mathrm{i} = \mathbb{E}\Big[\frac{\partial}{\partial \boldsymbol{\theta}_j} \log \frac{f_U(U, \mathbf{Z}; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\Big| \delta = 1\Big] = \int_{\mathbf{z}} \int_0^\infty \int_0^\infty \frac{\partial}{\partial \boldsymbol{\theta}_j} \log \frac{f_U(t+r, \mathbf{z}; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\ dP(t, r, \mathbf{z} | \delta = 1)\,. \tag{4.2.7}$$

Thus

$$
\begin{aligned}
\mathrm{i} &= \frac{1}{\varsigma}\int_0^\infty \int_0^\infty \frac{\partial}{\partial \boldsymbol{\theta}_j} \log \frac{f_U(u, \mathbf{z}; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} S_C(r) \frac{f_U(t+r, \mathbf{z}; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} dt dr d\mathbf{z} \\
&= \frac{1}{\varsigma}\int_{\mathbf{z}}\int_0^\infty \int_0^\infty \frac{\partial}{\partial \boldsymbol{\theta}_j} \Big(S_C(r) \frac{f_U(t+r, \mathbf{z}; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} \Big) dt dr d\mathbf{z} \\
&= 0,
\end{aligned}
$$

as

$$
\int_{\mathbf{z}}\int_0^\infty \int_0^\infty S_C(r) \frac{f(t+r, \mathbf{z}; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} dt dr d\mathbf{z} = \int_{\{t,r,\mathbf{z}\}} dP(t, r, \mathbf{z}, \delta = 1) = 1.
$$

and

$$
\begin{aligned}
\mathrm{ii} &= \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}_j} \log \Big(\int_{\omega \geq V} \frac{f_U(\omega, \mathbf{z}; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} d\omega \Big) \Big| \delta = 0 \right] \\
&= \int_{\mathbf{z}}\int_0^\infty \int_0^\infty \frac{\partial}{\partial \boldsymbol{\theta}_j} \log \Big(\int_{\omega \geq v} \frac{f_U(\omega, \mathbf{z}; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} d\omega \Big) \, dP(t, c, \mathbf{z} | \delta = 0).
\end{aligned} \tag{4.2.8}
$$

where

$$
\begin{aligned}
dP(x,\mathbf{z};\delta=0) &= P\big(T\in(t,t+dt], R\wedge C\in(c,c+dc], \mathbf{z}\in B_{d\mathbf{z}}(\mathbf{z}), R>T\big)\\
&= P\big(T\in(t,t+dt], C\in(c,c+dc], \mathbf{z}\in B_{d\mathbf{z}}(\mathbf{z}), R>C\big)\\
&= P\big(T\in(t,t+dt], \mathbf{z}\in B_{d\mathbf{z}}(\mathbf{z}), R>c\big) f_C(c)\ dc\\
&= \Big[\int_t^\infty f_{T,R,\mathbf{Z}}(t,r,\mathbf{z})dr\Big] f_C(c)\ dcdtd\mathbf{z}\\
&= f_C(c)\int_t^\infty \frac{f_U(t+r,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}dr\ dtdcd\mathbf{z}\\
&= f_C(c)\int_{t+c}^\infty \frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}d\omega\ dtdcd\mathbf{z}\\
&= f_C(c)\int_{\omega\geq v} \frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}d\omega\ dtdcd\mathbf{z}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\text{ii} &= \int_{\mathbf{z}}\int_0^\infty\int_0^\infty \Big[\frac{\partial}{\partial\boldsymbol{\theta}_j}\log\Big(\int_{\omega\geq v}\frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}d\omega\Big)\Big]\int_{\omega\geq v}\frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}d\omega\ f_C(c)dtdcd\mathbf{z}\\
&= \int_{\mathbf{z}}\int_0^\infty\int_0^\infty \frac{\frac{\partial}{\partial\boldsymbol{\theta}_j}\int_{\omega\geq v}\frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}d\omega}{\int_{\omega\geq v}\frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}d\omega}\int_{\omega\geq v}\frac{f_U(\omega;\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}d\omega f_C(c)dtdcd\mathbf{z}\\
&= \int_{\mathbf{z}}\int_0^\infty\int_0^\infty \frac{\partial}{\partial\boldsymbol{\theta}_j}\int_{\omega\geq v}\frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}d\omega dtdcd\mathbf{z}\\
&= 0
\end{aligned}
$$

as

$$
\int_{v,c,\mathbf{z}}\Big[\int_{\omega\geq v}\frac{f_U(\omega,\mathbf{z};\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}d\omega f_C(c)\Big]dtdcd\mathbf{z}=1
$$

where $v=t+c$.

Next we find $\mathrm{Var}(\delta_i U_i^T(\theta_0))$, where the variance is to be interpreted as the vector of component-wise variances of $U_i^T(\theta)$. We first note that

$$\mathrm{Var}\Big(\delta_i \frac{\partial}{\partial \boldsymbol{\theta}_j} \log g_{\boldsymbol{\theta}_0}(U_i, \mathbf{Z}_i)\Big) = \mathbb{E}\Big[\delta_i^2 \Big(\frac{\partial}{\partial \boldsymbol{\theta}_j} \log g_{\boldsymbol{\theta}_0}(U_i, \mathbf{Z}_i)\Big)^2\Big]$$

as $\mathbb{E}[\delta_i \frac{\partial}{\partial \boldsymbol{\theta}_j} \log g_{\boldsymbol{\theta}_0}(U_i, \mathbf{Z})] = 0$. Therefore,

$$\begin{aligned}
\mathrm{Var}\Big(\delta_i \frac{\partial}{\partial \boldsymbol{\theta}_j} \log g_{\boldsymbol{\theta}_0}(U_i, \mathbf{Z}_i)\Big) &= \mathbb{E}\Big[\delta_i \Big[\frac{\partial}{\partial \boldsymbol{\theta}_j} \log g_{\boldsymbol{\theta}_0}(U_i, \mathbf{Z}_i)\Big]^2\Big] \\
&= \mathbb{E}\Big[\Big[\frac{\partial}{\partial \boldsymbol{\theta}_j} \log g_{\boldsymbol{\theta}_0}(U_i, \mathbf{Z}_i)\Big]^2 | \delta_i = 1\Big] P(\delta_i = 1) \\
&= -\mathbb{E}\Big\{\Big[\frac{\partial^2}{\partial \boldsymbol{\theta}_j^2} \log g_{\boldsymbol{\theta}_0}(U_i, \mathbf{Z}_i) | \delta_i = 1\Big]\Big\} P(\delta_i = 1) \\
&= -\mathbb{E}\Big[\delta_i \frac{\partial^2}{\partial \boldsymbol{\theta}_j^2} \log g_{\boldsymbol{\theta}_0}(U_i, \mathbf{Z}_i)\Big].
\end{aligned}$$

The above equation and the strong law of large numbers imply that

$$\frac{1}{n} \sum_{i=1}^{n} \delta_i \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{U}_i^T(\boldsymbol{\theta}_0) \xrightarrow{a.s.} \mathrm{Var}(\delta \mathcal{U}^T(\boldsymbol{\theta}_0))$$

On the other hand,

$$\mathrm{Cov}\Big(\delta \frac{\partial}{\partial \boldsymbol{\theta}_j} \log g_{\boldsymbol{\theta}}(U, \mathbf{Z}), (1-\delta) \frac{\partial}{\partial \boldsymbol{\theta}_l} \log h_{\boldsymbol{\theta}}(V, \mathbf{Z})\Big) = 0,$$

Thus as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n}\Big[\delta_i \frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{U}_i^T(\boldsymbol{\theta}_0) + (1-\delta_i)\frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{V}_i^T(\boldsymbol{\theta}_0)\Big] \to \mathrm{Var}\Big[\delta\mathcal{U}^T(\boldsymbol{\theta}_0) + (1-\delta)\mathcal{V}^T(\boldsymbol{\theta}_0)\Big]. \tag{4.2.9}$$

Chapter 5

Covariates and length-biased sampling

With the asymptotic properties of the MLEs established, the focus of this chapter is the development of parametric models that take into account length-biased lifetimes with covariates, and the expression of the likelihood for those models.

5.1 Regression in survival analysis

Before looking at specific models, a few more preliminaries are needed. When it comes to modeling lifetimes, the objects of interest differ from more common statistical settings. For example, as mean lifetimes tend to be influenced by long living individuals, median lifetimes are more likely to be used as a measure of average. Also, the distribution function $F(x)$, gives the probability of living *at most* up to time x, which is of little use to patients, doctors and governments alike. This is why the survival function $S(x)$, giving the probability of living for at least x amount of time, is a much more meaningful measure to consider.

One particular notion that is central to survival analysis is the concept of hazard, which one can express as an instant rate of failure (or whichever is the event under study), denoted by $\lambda(x)$. It is formally defined by:

$$\lambda(x) = \lim_{dx \to 0} \frac{P\{x \leq X < x + dx | X \geq x\}}{dx}. \tag{5.1.1}$$

One can then extend this idea into the notion of cumulative hazard:

$$\Lambda(x) = \int_0^x \lambda(s)\, ds. \tag{5.1.2}$$

A number of relationships can be established, the most useful for this thesis being the following:

$$S(x) = \exp\Big(- \Lambda(x)\Big) = \exp\Big(- \int_0^x \lambda(s)\, ds\Big), \tag{5.1.3}$$

and

$$f(x) = \lambda(x) S(x). \tag{5.1.4}$$

Note that, as one can write the mean through integration of the survival function, it can also be written using only hazard:

$$\begin{aligned}\mu &= \int_0^\infty S(x)\,dx \\ &= \int_0^\infty \exp\big(-\Lambda(x)\big)\,dx \\ &= \int_0^\infty \exp\Big(-\int_0^x \lambda(s)\,ds\Big)\,dx\end{aligned} \tag{5.1.5}$$

One can use 5.1.4 to rewrite the likelihood in terms of the hazard and survival function:

$$\mathcal{L}_I(\boldsymbol{\theta}) = \prod_{i=1}^n \lambda^{\delta_i}(x_i|\mathbf{z}_i;\boldsymbol{\theta})\frac{S(x_i|\mathbf{z}_i;\boldsymbol{\theta})}{\mu(\mathbf{z}_i;\boldsymbol{\theta})}, \tag{5.1.6}$$

$$\mathcal{L}_J(\boldsymbol{\theta}) = \prod_{i=1}^n \lambda^{\delta_i}(x_i|\mathbf{z}_i;\boldsymbol{\theta})\frac{S(x_i|\mathbf{z}_i;\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}. \tag{5.1.7}$$

The number of commonly used parametric models for which $\lambda(x|\mathbf{z};\boldsymbol{\theta})$, $S(x|\mathbf{z};\boldsymbol{\theta})$ and $\mu(\mathbf{z};\boldsymbol{\theta})$ have a closed form is limited. They include Weibull and Pareto models which will be discussed in the next sections.

One of the basic regression models in survival analysis is the proportional hazards model. Modeling hazard has advantages over working directly with the survival function, as the hazard function can assume a variety of shapes while the survival function is inevitably monotone decreasing and bouded. This means hazard is more flexible and its different forms are easier intuitively conceptualize and interpret. The proportional hazard model makes the assumption that covariates change the hazard only through multiplication by a constant factor. Mathematically,

$$\lambda(x|\mathbf{z}) = \lambda_0(x)e^{\boldsymbol{\beta}'\mathbf{z}}. \tag{5.1.8}$$

While very popular due to the use of Cox's model which is semiparametric and does not impose any structure on the baseline hazard, one should always check if this assumption of proportionality of hazard holds before employing such a model.

When assuming a proportional hazard model for the unbiased population, in the length-biased environment things are much different:

$$\begin{aligned}
\lambda_{LB}(x|\mathbf{z}) &= \frac{f_{LB}(x|\mathbf{z})}{S_{LB}(x|\mathbf{z})} \\
&= \frac{\frac{xf_U(x|\mathbf{z})}{\mu(\mathbf{z})}}{\int_x^\infty \frac{sf_U(s|\mathbf{z})}{\mu(\mathbf{z})}\,ds} \\
&= \frac{x\lambda_0(x)\exp\big(-e^{\boldsymbol{\beta}'\mathbf{z}}\Lambda_0(x)\big)}{\int_x^\infty s\lambda_0(s)\exp\big(-e^{\boldsymbol{\beta}'\mathbf{z}}\Lambda_0(s)\big)\,ds}
\end{aligned} \tag{5.1.9}$$

Fortunately, as one can write the likelihood in terms of the unbiased distribution (and related quantities), this does not pose a problem in terms of implementation.

The other basic regression model in survival data analysis is the accelerated failure-time model (AFT). It can be expressed in terms of the survival function:

$$S(x|\mathbf{z};\boldsymbol{\beta}) = S_0(xe^{-\boldsymbol{\beta}'\mathbf{z}}), \tag{5.1.10}$$

where $S_0(x)$ is the baseline survival function when the covariates are all 0. This model assumes that covariates modify the time-scale of the event of interest. Though it can be seen as an extension of linear regression for positive random variables through a logarithm transformation, the number of parametric error distributions it can accommodate in closed form is limited.

If one assumes an accelerated failure-time on an unbiased population, the resulting length-biased population also follows an accelerated failure-time model with the same covariate effects, if taken conditionally. First, we have to go to the density function:

$$\begin{aligned} f_U(x|\mathbf{z};\boldsymbol{\beta}) &= -\frac{dS(x|\mathbf{z};\boldsymbol{\beta})}{dx} = -\frac{dS_0(xe^{-\boldsymbol{\beta}'\mathbf{z}})}{dx} \\ &= \big(-S_0'(xe^{-\boldsymbol{\beta}'\mathbf{z}})\big)e^{-\boldsymbol{\beta}'\mathbf{z}} \\ &= f_{U,0}(xe^{-\boldsymbol{\beta}'\mathbf{z}})e^{-\boldsymbol{\beta}'\mathbf{z}}, \end{aligned} \tag{5.1.11}$$

where $f_{U,0}(x) = -dS_0(x)/dx$. For all AFT models, $\mu(\mathbf{z};\boldsymbol{\beta})$ can be expressed as

$$\begin{aligned} \mu(\mathbf{z};\boldsymbol{\beta}) &= \int_0^\infty S(x|\mathbf{z};\boldsymbol{\beta})\,dx \\ &= \int_0^\infty S_0(xe^{-\boldsymbol{\beta}'\mathbf{z}})\,dx \\ &= e^{\boldsymbol{\beta}'\mathbf{z}}\int_0^\infty S_0(w)dw \\ &= e^{\boldsymbol{\beta}'\mathbf{z}}\mu(0), \end{aligned} \tag{5.1.12}$$

where $\mu(0)$ is the baseline mean lifetime. So the length-biased density of an

AFT model is given by:

$$
\begin{aligned}
f_{LB}(x|\mathbf{z};\boldsymbol{\beta}) &= \frac{x f_U(x|\mathbf{z};\boldsymbol{\beta})}{\mu(\mathbf{z};\boldsymbol{\beta})} \\
&= \frac{x e^{-\boldsymbol{\beta}'\mathbf{z}} f_{U,0}(x e^{-\boldsymbol{\beta}'\mathbf{z}})}{e^{\boldsymbol{\beta}'\mathbf{z}}\mu(0)} \\
&= \left(\frac{x e^{-\boldsymbol{\beta}'\mathbf{z}} f_{U,0}(x e^{-\boldsymbol{\beta}'\mathbf{z}})}{\mu(0)}\right) e^{-\boldsymbol{\beta}'\mathbf{z}} \\
&= f_{LB,0}(x e^{-\boldsymbol{\beta}'\mathbf{z}}) e^{-\boldsymbol{\beta}'\mathbf{z}},
\end{aligned}
\tag{5.1.13}
$$

where $f_{LB,0}(x) = \frac{x f_{U,0}(x)}{\mu(0)}$. Hence

$$
S_{LB}(x|\mathbf{z};\boldsymbol{\beta}) = S_{LB,0}(x e^{-\boldsymbol{\beta}'\mathbf{z}}) \tag{5.1.14}
$$

as desired.

Because of this apparent equivalence between unbiased and length-biased AFT models, one might be tempted to simply apply the methods available for incident cases (right-censored with no truncation) to the length-biased distribution and estimate the covariate effects. This would be wrong for two reasons. First these methods rely on the assumption that censoring is uninformative, which is not the case with prevalent cohort. Secondly it would still ignore information about the covariate effects contained in the sampling distribution of the covariates. Speaking of which, following 5.1.12:

$$
\begin{aligned}
f_B(\mathbf{z};\boldsymbol{\beta}) &= \frac{\mu(\mathbf{z};\boldsymbol{\beta}) f_{\mathbf{Z}}(\mathbf{z})}{\int_{\mathbf{z}} \mu(\mathbf{z};\boldsymbol{\beta}) f_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z}} \\
&= \frac{e^{\boldsymbol{\beta}'\mathbf{z}} \mu(0) f_{\mathbf{Z}}(\mathbf{z})}{\int_{\mathbf{z}} e^{\boldsymbol{\beta}'\mathbf{z}} \mu(0) f_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z}} \\
&= \frac{e^{\boldsymbol{\beta}'\mathbf{z}} f_{\mathbf{Z}}(\mathbf{z})}{\mathbb{E}(e^{\boldsymbol{\beta}'\mathbf{Z}})}
\end{aligned}
\tag{5.1.15}
$$

holds for all AFT models, which means the sampling distribution of the covariates depends only on the covariate effects regardless of the lifetime distribution. The implications of this compared to proportional hazard models are discussed in the next section.

5.2 Length-biased sampling and Weibull models

Because of its wide use and versatility, it is appropriate to consider the Weibull distribution. Recall that, for positive x, the Weibull density is given by

$$
f_U(x) = \alpha\lambda^{\alpha} x^{\alpha-1} \exp\big(-(\lambda x)^{\alpha}\big). \tag{5.2.1}
$$

One may note that the mean of a Weibull(α, λ) parametrized as in 5.2.1 is

$$\mu(\alpha, \lambda) = \frac{\Gamma(1 + 1/\alpha)}{\lambda}. \tag{5.2.2}$$

One can express a Weibull model as a proportional hazards model:

$$\lambda(x|\mathbf{z}; \alpha, \lambda, \boldsymbol{\beta}) = e^{\boldsymbol{\beta}'\mathbf{z}} \alpha \lambda^{\alpha} x^{\alpha-1} = \alpha\big(e^{\frac{\boldsymbol{\beta}'\mathbf{z}}{\alpha}} \lambda\big)^{\alpha} x^{\alpha-1}. \tag{5.2.3}$$

Hence, when given a vector of covariates $\mathbf{z}$, this is the hazard function of a Weibull$(\alpha, e^{\frac{\boldsymbol{\beta}'\mathbf{z}}{\alpha}} \lambda)$, with mean

$$\mu(\mathbf{z}; \alpha, \lambda, \boldsymbol{\beta}) = \frac{\exp\big(-\frac{\boldsymbol{\beta}'\mathbf{z}}{\alpha}\big)\Gamma(1 + 1/\alpha)}{\lambda}. \tag{5.2.4}$$

Assuming a Weibull model in the incident population (the population of interest) does not lead to a Weibull distribution for a length-biased prevalent cohort. It should be mentioned that the Weibull distribution is a special case of the Generalized Gamma distribution (henceforth denoted GG) with density given by

$$f(x) = \frac{\lambda \alpha}{\Gamma(\gamma)} (\lambda x)^{\alpha\gamma-1} \exp\big(-(\lambda x)^{\alpha}\big). \tag{5.2.5}$$

Correa & Wolfson (1999) demonstrate that the length-biased distribution of a GG$(\alpha, \lambda, \gamma)$ is a GG$(\alpha, \lambda, \gamma + 1/\alpha)$.

The marginal sampling density of the covariates, from an unbiased Weibull population, can be expressed as

$$f_B(\mathbf{z};\alpha,\lambda,\boldsymbol{\beta}) = \frac{\exp\left(-\frac{\boldsymbol{\beta}'\mathbf{z}}{\alpha}\right)f_{\mathbf{Z}}(\mathbf{z})}{\displaystyle\int_{\mathbf{z}} \exp\left(-\frac{\boldsymbol{\beta}'\mathbf{z}}{\alpha}\right)f_{\mathbf{Z}}(\mathbf{z})d\mathbf{z}}, \tag{5.2.6}$$

where, when $\mathbf{Z}$ is univariate, the denominator reduces to $M_Z(-\beta/\alpha)$, the moment generating function of Z provided it exists. Note that this sampling density not only depend on the covariate effects but it also involves α, the shape parameter of the Weibull density of the lifetimes. In particular, as $f_B(\mathbf{z})$ depends only on the ratio of $\boldsymbol{\beta}$ and α, it is possible that this might lead to some identifiability problems.

We can now write the likelihoods for Weibull data, starting with the one that ignores the informativeness of the covariates as it has an explicit formula:

$$\begin{aligned}\mathcal{L}_I\Big(((x_1,\mathbf{z}_1),\ldots,(x_n,\mathbf{z}_n));\alpha,\lambda,\boldsymbol{\beta}\Big)&\\ &= \prod_{i=1}^{n}\frac{1}{\mu(\mathbf{z}_i;\alpha,\lambda,\boldsymbol{\beta})}\left(e^{\boldsymbol{\beta}'\mathbf{z}_i}\alpha\lambda^\alpha x_i^{\alpha-1}\right)^{\delta_i}\exp\left(-e^{\boldsymbol{\beta}'\mathbf{z}_i}(\lambda x_i)^\alpha\right)\\ &= \prod_{i=1}^{n}\frac{e^{\frac{\boldsymbol{\beta}'\mathbf{z}_i}{\alpha}}\lambda}{\Gamma(1+\frac{1}{\alpha})}\left(e^{\boldsymbol{\beta}'\mathbf{z}_i}\alpha\lambda^\alpha x_i^{\alpha-1}\right)^{\delta_i}\exp\left(-e^{\boldsymbol{\beta}'\mathbf{z}_i}(\lambda x_i)^\alpha\right)\end{aligned} \tag{5.2.7}$$

The correct likelihood is not so elegant:

$$\mathcal{L}_J\Big(\big((x_1, \mathbf{z}_1), \ldots, (x_n, \mathbf{z}_n)\big); \alpha, \lambda, \boldsymbol{\beta}\Big)$$
$$= \prod_{i=1}^{n} \frac{1}{\mu(\alpha, \lambda, \boldsymbol{\beta})} \big(e^{\boldsymbol{\beta}'\mathbf{z}_i} \alpha \lambda^\alpha x_i^{\alpha-1}\big)^{\delta_i} \exp\big(- e^{\boldsymbol{\beta}'\mathbf{z}_i} (\lambda x_i)^\alpha\big) \tag{5.2.8}$$

The difficulty here is that $\mu(\alpha, \lambda, \boldsymbol{\beta})$ does not have an explicit formula unless one knows $f_{\mathbf{Z}}(\mathbf{z})$ *a priori*, which not often the case in real applications. When $f_{\mathbf{Z}}(\mathbf{z})$ is unknown, there are at least two ways to approach $\mu(\alpha, \lambda, \boldsymbol{\beta})$ in $\mathcal{L}_J$. The first way is to use 3.4.3 and obtain

$$\frac{1}{\mu(\alpha, \lambda, \boldsymbol{\beta})} = \frac{\lambda}{\Gamma(1 + \frac{1}{\alpha})} \int_{\mathbf{z}} e^{\frac{\boldsymbol{\beta}'\mathbf{z}}{\alpha}} f_B(\mathbf{z}) d\mathbf{z}. \tag{5.2.9}$$

The density $f_B(\mathbf{z})$ has a natural nonparametric estimate through the sample distribution of the covariates. Unfortunately this cannot be used for implementation, as numerically, $\sum_{i=1}^{n} \frac{1}{n} e^{\frac{\boldsymbol{\beta}'\mathbf{z}_i}{\alpha}}$ will explode for positive $\boldsymbol{\beta}'\mathbf{z}$ as $\alpha \to 0$. The second approach to express μ is through 3.4.1:

$$\mu(\alpha, \lambda, \boldsymbol{\beta}) = \frac{\Gamma(1 + \frac{1}{\alpha})}{\lambda} \int_{\mathbf{z}} \exp\Big(- \frac{\boldsymbol{\beta}'\mathbf{z}}{\alpha}\Big) f_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z}. \tag{5.2.10}$$

The expression now involves finding the unbiased expectation of $\exp(-\frac{\boldsymbol{\beta}'\mathbf{z}}{\alpha})$, which will be numerically more stable, provided one has some knowledge of $f_{\mathbf{Z}}(\mathbf{z})$. In this form, one might want to simply assume that $f_{\mathbf{Z}}(\mathbf{z})$ is known under reasonable circumstances but in most cases it will have to be estimated.

To that end, a new method to estimate $f_{\mathbf{Z}}(\mathbf{z})$ nonparametrically is proposed in the next chapter.

One can also express a Weibull model as an AFT model. Let $S_0(x)$ be the survival function of a Weibull(α, λ), then

$$S(x|\mathbf{z}) = S_0(xe^{-\boldsymbol{\beta}'\mathbf{z}}) = \exp\big(-(\lambda x e^{-\boldsymbol{\beta}'\mathbf{z}})^{\alpha}\big) = \exp\big(-(\lambda e^{-\boldsymbol{\beta}'\mathbf{z}}x)^{\alpha}\big), \quad (5.2.11)$$

which is the survival function of a Weibull$(\alpha, \lambda e^{-\boldsymbol{\beta}'\mathbf{z}})$.

As observed in the previous section, the sampling density of the covariate does not depend on α:

$$\mu(\mathbf{z}; \alpha, \lambda, \boldsymbol{\beta}) = \frac{\Gamma(1+1/\alpha)}{\lambda e^{-\boldsymbol{\beta}'\mathbf{z}}} = e^{\boldsymbol{\beta}'\mathbf{z}}\left(\frac{\Gamma(1+1/\alpha)}{\lambda}\right) = e^{\boldsymbol{\beta}'\mathbf{z}}\mu(0). \qquad (5.2.12)$$

As

$$f_B(\mathbf{z}; \boldsymbol{\theta}) = \frac{e^{\boldsymbol{\beta}'\mathbf{z}}\mu(0) f_{\mathbf{Z}}(\mathbf{z})}{\int_{\mathbf{z}} e^{\boldsymbol{\beta}'\mathbf{z}}\mu(0) f_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z}} = \frac{e^{\boldsymbol{\beta}'\mathbf{z}} f_{\mathbf{Z}}(\mathbf{z})}{\mathbb{E}(e^{\boldsymbol{\beta}'\mathbf{Z}})}, \qquad (5.2.13)$$

only depends on $\boldsymbol{\beta}$, the covariate effects, the foreseen difficulties in numerically maximizing the likelihood for proportional hazard models may not occur for AFT models. Note how the overall mean is proportional to $\mathbb{E}(e^{\boldsymbol{\beta}'\mathbf{Z}})$ and $\mu(\alpha, \lambda, \boldsymbol{\beta}')^{-1} \propto \mathbb{E}_B(e^{-\boldsymbol{\beta}'\mathbf{Z}})$, the interpretation of the parameters for covariate effect is the opposite of the proportional hazard case.

5.3 Length-biased sampling and Pareto models

While perhaps less widely used than Weibull models, the Pareto distribution also has some interesting properties when it comes to regression, both with AFT and proportional hazard models. Using the same set of parameters as before (which might not be standard notation), the probability density function is given by this:

$$f_U(x;\alpha,\lambda) = \frac{\alpha\lambda^\alpha}{x^{\alpha+1}}, \tag{5.3.1}$$

where $\alpha > 0$, $\lambda > 0$ and $x > \lambda$. The mean exists only if $\alpha > 1$:

$$\mu(\alpha,\lambda) = \frac{\alpha\lambda}{\alpha - 1}. \tag{5.3.2}$$

Interestingly, the length-biased distribution of a Pareto distribution is also Pareto with parameters $(\alpha - 1, \lambda)$:

$$f_{LB}(x;\alpha,\lambda) = \frac{x f_U(x)}{\mu} = x\frac{\alpha\lambda^\alpha}{x^{\alpha+1}}\frac{\alpha-1}{\alpha\lambda} = \frac{(\alpha-1)\lambda^{\alpha-1}}{x^\alpha} \tag{5.3.3}$$

A Pareto model can be expressed as a proportional hazards model:

$$\lambda(x|\mathbf{z};\alpha,\boldsymbol{\beta}) = e^{\boldsymbol{\beta}'\mathbf{z}}\lambda_0(x;\alpha) = \frac{e^{\boldsymbol{\beta}'\mathbf{z}}\alpha}{x} \tag{5.3.4}$$

which is the hazard of a Pareto($e^{\boldsymbol{\beta}'\mathbf{z}}\alpha, \lambda$). In this case the mean conditional on the covariates $\mathbf{z}$ is:

$$\mu(\mathbf{z}; \alpha, \lambda, \boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta}'\mathbf{z}}\alpha\lambda}{e^{\boldsymbol{\beta}'\mathbf{z}}\alpha - 1} = \frac{\alpha\lambda}{\alpha - e^{-\boldsymbol{\beta}'\mathbf{z}}}, \tag{5.3.5}$$

for $e^{\boldsymbol{\beta}'\mathbf{z}}\alpha > 1$. The sampling density of the covariates depends on $\boldsymbol{\beta}$ and α:

$$f_B(\mathbf{z}; \alpha, \boldsymbol{\beta}) = \frac{(\alpha - e^{-\boldsymbol{\beta}'\mathbf{z}})^{-1} f_{\mathbf{Z}}(\mathbf{z})}{\mathbb{E}\big((\alpha - e^{-\boldsymbol{\beta}'\mathbf{z}})^{-1}\big)} \tag{5.3.6}$$

An AFT model based on a Pareto distribution yields a Pareto model:

$$S_0(x; \alpha, \lambda) = \frac{\lambda^\alpha}{x^\alpha}, \tag{5.3.7}$$

$$S(x|\mathbf{z}) = S_0(xe^{-\boldsymbol{\beta}'\mathbf{z}}) = \frac{\lambda^\alpha}{(xe^{-\boldsymbol{\beta}'\mathbf{z}})^\alpha} = \frac{(e^{\boldsymbol{\beta}'\mathbf{z}}\lambda)^\alpha}{x^\alpha}, \tag{5.3.8}$$

which is the survival function of a Pareto($\alpha, e^{\boldsymbol{\beta}'\mathbf{z}}\lambda$). Consequently its mean is

$$\mu(\mathbf{z}; \alpha, \lambda, \boldsymbol{\beta}) = \frac{\alpha e^{\boldsymbol{\beta}'\mathbf{z}}\lambda}{\alpha - 1} = e^{\boldsymbol{\beta}'\mathbf{z}}\mu(\alpha, \lambda), \tag{5.3.9}$$

as predicted by 5.1.12.

Chapter 6

Algorithms

6.1 Nonparametric estimation of $f_{\mathbf{Z}}(\mathbf{z})$

Before discussing the simulation of length-biased data with covariates, we introduce an algorithm to find an estimate for the unbiased distribution of the covariates, which is necessary for the implementation of $\mathcal{L}_J(\boldsymbol{\theta})$. Consider Vardi's Problem A, where there are M complete independent observations $U_1, \ldots, U_M$ with distribution G and N independent incomplete observations $V_1, \ldots, V_N$, in the sense that $V_i = W_i C_i$, where W_i has distribution G and C_i is independently drawn from a uniform distribution on (0,1). The likelihood for the observations in Problem A is given by:

$$\mathcal{L}(G) = \prod_{i=1}^{M} G(du_i) \prod_{i=1}^{N} \int_{s \geq v_i} \frac{1}{s} G(ds). \tag{6.1.1}$$

Note that, in this problem, the number of censored and uncensored observations is known *a priori*, and that the setting appears completely different than prevalent cohort data. Nevertheless, once we rewrite 3.2.2 in terms of 6.1.1, we can proceed to the nonparametric likelihood. While the data we are considering consist of observations of the form $(x_i, \mathbf{z}_i, \delta_i)$, one may ignore the covariates for the time being, and split the sample in two groups. The pairs $(x_i, 1)$ are the uncensored observations, so they can be relabeled as u_i's, and the $(x_i, 0)$ pairs will be the v_i's. We can write $n = M + N$, where M and N fit Vardi's original problem. Since the likelihood considers the observations, in particular the u_i's, as constants, the distribution that maximizes 3.2.2 also maximizes

$$\mathcal{L}(f_U) \propto \prod_{i=1}^{M} \frac{u_i f(u_i)}{\mu} \prod_{i=1}^{N} \int_{s \geq v_i} \frac{f_U(s)}{\mu} ds \tag{6.1.2}$$

Letting G be the length-biased distribution of the data, in other words,

$$G(du) = \frac{u f_U(u)}{\mu} du = F_{LB}(du), \tag{6.1.3}$$

the equivalence between 3.2.2 and 6.1.1 becomes evident.

As shown by Vardi, the likelihood 6.1.1 can be maximized nonparametrically, by assigning positive weights only on the observed u_i's and v_i's, because giving mass anywhere else would decrease the likelihood. Let $x_1^* < x_2^* < \cdots < x_m^*$ represent all the sorted distinct times (both complete and incomplete). Note that $m \leq M + N$, with equality if and only if there are no ties in the observed times (which in theory occurs with probability 1 when the underlying distribution is absolutely continuous, though in real applications we may observe ties). Let ξ_j be the number of complete observations at x_j^* and ζ_j be the number of censored observations at x_j, define $\mathbf{p}^* = (p_1^*, \ldots, p_m^*)'$ to be a probability vector, where $p_j^* = P(x_j^*) = G(dx_j^*)$, then 6.1.1 can be expressed as:

$$\mathcal{L}(\mathbf{p}^*) = \prod_{i=1}^{h} p_j^{*\xi_j} \Big(\sum_{k=j}^{h} \frac{1}{x_k^*} p_k^* \Big)^{\zeta_j} \qquad (6.1.4)$$

Since we have incomplete data (which contains both complete observations through the observed failure times and incomplete observations in the censored subjects), but the form of the complete data is known, Vardi notes that the EM algorithm comes as a natural solution to maximizing the likelihood in 6.1.4. If all the observations were uncensored (i.e. complete data), the maximum likelihood distribution would be the empirical distribution of the data, with mass proportional to the number of observations at each time point. Given a previous estimate for $\mathbf{p}^*$, say $\mathbf{p}^{*,old}$, the EM algorithm assigns to each $p_j^{*,new}$ the expected conditional number of complete-data-observations at x_j^* given the observed data and the previous probability vector $\mathbf{p}^{*,old}$, divided by the total number of observations, so that $\mathbf{p}^{*,new}$ is also a probability vector. The mathematical formulation can be written in two points.

Algorithm 1.

- *Start with an initial guess probability vector* $\mathbf{p}^{*,old}$ *with positive mass at each time point.*

- *Update* $\mathbf{p}^*$ *at each time point using*

$$p_j^{*new} = \frac{1}{M+N} E\Big(\sum_{i=1}^{M} I[u_i = x_j^*] + \sum_{i=1}^{N} I[v_i = x_j^*]\Big|(u_1, \ldots, u_M, v_1, \ldots, v_N), \mathbf{p}^{old}\Big)$$
$$= \frac{1}{M+N}\Big(\xi_j + x_j^{*-1} p_j^{old} \sum_{k=1}^{j} \zeta_k \big(\sum_{i=k}^{h} x_i^{*-1} p_i^{*,old}\big)^{-1}\Big). \tag{6.1.5}$$

One might note that if none of the observations are censored, all ζ_i's are 0, $\hat{\mathbf{p}}^*$ is fixed and there is no iteration. By virtue of the EM algorithm, 6.1.5 will converge to $\hat{\mathbf{p}}^*$, the unique maximizer of 6.1.4, which in turn yields a consistent estimate for the length-biased distribution. To recover the unbiased distribution, one needs to readjust the probabilities in $\hat{\mathbf{p}}^*$ using the inverse length-bias transformation:

$$\hat{p}_{U,j}^* = \frac{\hat{p}_j^*/x_j^*}{\sum_{i=1}^{h} \hat{p}_i^*/x_i^*}. \tag{6.1.6}$$

In the same way 3.2.2 was extended to 3.2.19, one can consider the observations as points in a space of dimension $1+k$, the positive real line joint with the space of the covariates. As such, Vardi's estimate can be taken as an estimate of the joint unbiased distribution of time and the covariates. It should be noted that, in the case of ties in time, the weight of each time point has to be spread over each distinct observation at that time point:

$$\hat{p}(x_i, \mathbf{z}_i) = \frac{\hat{p}^*_U(x^*_j = x_i)}{\xi_j + \zeta_j} \tag{6.1.7}$$

One can obtain an estimate for the unbiased distribution of the covariates by taking the right marginal:

$$\hat{p}_{\mathbf{Z}}(\mathbf{z}) = \sum_{j \text{ s.t. } \mathbf{z}_j = \mathbf{z}} \hat{p}_U(x_j, \mathbf{z}_j). \tag{6.1.8}$$

6.2 Simulating length-biased data with covariates

In this section, we discuss how to simulate length-biased data with covariates when the times come from an unbiased mixture of Weibull populations. As explained in Section 6, a proportional hazard model on a Weibull distribution yields a Weibull distribution for given a covariate $\mathbf{z}$. When covariates are discrete, this mean that the incident population comes from a finite or countable set of distinct Weibull distribution with the same shape parameter α but a rate parameter $\lambda(\mathbf{z})$ which depends on a baseline rate λ, the covariate effects $\boldsymbol{\beta}$ and the common α. We can look at the joint density in mathematical terms:

$$f_U(x, \mathbf{z}; \alpha, \lambda, \boldsymbol{\beta}) = f_U(x|\mathbf{z}; \alpha, \lambda, \boldsymbol{\beta}) p_{\mathbf{Z}}(\mathbf{z}). \tag{6.2.1}$$

Now this is clearly a mixture distribution but as the covariates are observed we have no difficulty ascertaining which observation belongs to which subpopulation as the mixture probabilities are given by the marginal covariate probabilities and contain no parameter of interest. When we sample with length-bias, we again have a mixture of distinct distributions (now GGs which can be expressed in terms of the unbiased Weibull distributions):

$$f_{LB}(x, \mathbf{z}; \alpha, \lambda, \boldsymbol{\beta}) = \frac{x f_U(x|\mathbf{z}; \alpha, \lambda, \boldsymbol{\beta})}{\mu(\mathbf{z}; \alpha, \lambda, \boldsymbol{\beta})} p_B(\mathbf{z}; \alpha, \lambda, \boldsymbol{\beta}). \tag{6.2.2}$$

Now the mixture probabilities are again just the marginal probabilities of the covariates but depend on the parameters that need to be estimated and the unbiased probability distribution of the covariates which now cannot be brushed aside as it influences the unbiased mean of the incident population. Note that the same principles hold with continuous covariates.

There are at least two ways to generate length-biased data with covariates, though in both cases, the issues of truncation and censoring are handled in the exact same manner. The first approach is to start by generating a very large set of observations from the unbiased population including covariates and then sample from this superset a group of observations of the desired sample size with weight proportional to the time variable of each observation. This approach has the advantage that it can be applied to any distribution, and involves generating observations from models for incident data. Simulating observations from proportional hazard models is discussed by Bender, Augustin and Blettner (2005). The disadvantage of this method is that it is computationally expensive and inefficient.

The second approach involves sampling directly from the length-biased population. Correa & Wolfson (1999) give a simple algorithm to generate length-biased data from an unbiased Weibull(α, λ) population.

Algorithm 2. *Let*

$$g(s) = s^{1/\alpha}/\lambda. \tag{6.2.3}$$

To generate the desired length-biased data requires two steps:

- *Generate W say, from a Gamma$(1 + 1/\alpha, 1)$.*
- *Set $Y = g(W)$, where g is from 6.2.3.*

Then Y is distributed according to a $GG(\alpha, \lambda, 1 + 1/\alpha)$ as required.

As we know the general form of the biased distribution of the covariates, the problem of simulation reduces to using the most efficient algorithm to generate from the marginal $f_B(\mathbf{z})$. From that point, the simulation of joint length-biased times and covariates, with uniform left-truncation and right-censoring can be performed in a straight-forward way.

Algorithm 3. *The samples are generated as follows:*

- *First, pick values for α, λ, and $\boldsymbol{\beta}$, fix the sample size to n.*
- *Generate n covariates $\mathbf{z}$, from 5.2.6.*

- *For each* $\mathbf{z}_i$, *generate a length-biased time* y_i *from the appropriate GG.*
- *For each* y_i, *generate the truncation time* t_i *from a* $U(0, y_i)$.
- *The residual lifetime for each observation, is set to* $r_i = y_i - t_i$.
- *Generate the censoring times* c_i *(or pick* c *and let* $c_i = c$ *for constant censoring time).*
- *Let* $x_i = t_i + r_i \wedge c_i$, *and set* $\delta_i = I[r_i < c_i]$.

The variable δ serves as an "event-time" indicator, that is $\delta_i = 1$ when x_i is an event-time, and $\delta_i = 0$ when x_i is a censored observation. For the purpose of our simulation, which is for the implementation of $\mathcal{L}_I$ and $\mathcal{L}_J$, it is not essential to keep track of the truncation times. The data have the form $(x_i, \mathbf{z}_i, \delta_i)$.

6.3 Semiparametric bootstrap

Here, we propose a new semiparametric bootstrap method to obtain confidence bands for the estimated survival function. The use of bootstrap methods has grown with the advent of computational statistics. As their concept is easy to understand and because they are versatile and usually not very hard to program, it is of little surprise that bootstrap methods have gained much notoriety in empirical research. Efron & Tibshirani (1986) and Efron (1987) studied how to use bootstrap as a mean for obtaining variance estimates and confidence intervals as we want to do here. Semiparametric bootstrap methods have been developed for simulating extreme order statistics (see Zelterman, 1993) and to obtain confidence intervals for the Hurst coefficient (Hall et al., 2000). The issue of accuracy in bootstrap simulations has been looked into by Andrews and Buchinsky (2000) who proposed a method to select the number of bootstrap repetitions in a variety of settings. In the context of lifetime data analysis, Efron (1981) offered bootstrap methods for censored but not truncated data, Bilker and Wang (1997) extended that work to nonparametric bootstrap of left-truncated and right-censored data (in this case, with a general, unspecified truncation scheme).

In our case, the truncation distribution is defined to be uniform under the stationarity assumption. Since we have a parametric model for the survival times, it is likely more efficient to bootstrap length-biased lifetimes parametrically while nonparametrically sampling the covariates and censoring times. Unlike the simulated samples from the previous section, one needs to have the observed times from onset to recruitment and residual lifetimes (or censoring), not just the time from onset until death (or loss to follow-up). The set up is similar to algorithm 3, but with $\boldsymbol{\theta}$ and $f_C(c)$ being estimated from the data instead of being known *a priori*. Note that, as the residual lifetimes R_i's and the residual censoring times C_i's are independent, one can use symmetry to view the C_i's as being right-censored by the R_i's, and obtain an estimate for $S_C(c)$ using the Kaplan-Meier estimator (Kaplan & Meier, 1958). Because of the relationship between $\boldsymbol{\theta}$ and $f_B(\mathbf{z};\boldsymbol{\theta})$, the distribution of the covariates is kept *fixed* in the bootstrap sample.

Algorithm 4. *To obtain a bootstrap sample of length-biased data with uniform left truncation, right censoring and covariates from a sample of size n:*

- *Obtain $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\lambda}, \hat{\boldsymbol{\beta}})$ from the original data using $\mathcal{L}_I$ or $\mathcal{L}_J$.*
- *Estimate $S_C(c)$ using the Kaplan-Meier estimator of the residual censoring times*
- *Generate n censoring times c_i's from $\widehat{S_C(c)}$.*
- *Reproduce n covariates $\mathbf{z}_i$ as they are distributed in the observed data.*
- *For each $\mathbf{z}_i$, generate a length-biased time y_i from the appropriate GG.*
- *For each y_i, generate the truncation time t_i from a $U(0, y_i)$.*
- *The residual lifetime for each observation is set to $r_i = y_i - t_i$.*
- *Let $x_i = t_i + r_i \wedge c_i$, and set $\delta_i = I[r_i < c_i]$.*

Generating covariates from the empirical distribution (hence nonparametric) $\widehat{F_B(\mathbf{z})}$ is problematic as it causes bias in the bootstrap estimates for $\boldsymbol{\theta}$. Alternatively, one could replace $\widehat{F_B(\mathbf{z})}$ through a combination of the algorithm of section 7 and 5.2.6 using $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$; that is, obtain a semiparametric estimate for the biased distribution of the covariates. While this takes into account the relationship between $\boldsymbol{\theta}$ and mean lifetime bias in the covariates, it also requires more work and is probably inefficient statistically, as it involves estimating the unbiased distribution of the covariates which is sensitive to the inverse length-bias transformation when $\boldsymbol{\beta}$ is small, thereby inducing spurious variation in the bootstrap samples.If the unbiased distribution of the covariates was known *a priori*, then a parametric estimate for $f_B(\mathbf{z})$ might be useful.

Finally, by repeating the procedure of Algorithm 4 a large number of times, one can obtain new bootstrap estimates for the parameters and get confidence bands for the estimated survival curve, as well as compare the efficiency of $\mathcal{L}_J$ over $\mathcal{L}_I$ for estimation of survival curve at each point in time.

Chapter 7

Applications

7.1 Simulation studies

To illustrate the behaviour of $\mathcal{L}_I$ and $\mathcal{L}_J$ under different parameter values and censoring schemes, the results of a number of simulations of three types are presented. In the first case, we assume $f_{\mathbf{Z}}(\mathbf{z})$ is known and the two likelihoods are tested on binary mixture data, i.e. the data come from a population consisting of two distinct subpopulations, identified by a factor Z which is either 0 or 1, with equal probability in the unbiased population. The biased distribution of the covariates is then:

$$P_B(Z = 1; \alpha, \beta) = \frac{e^{-\beta/\alpha}}{1 + e^{-\beta/\alpha}}. \tag{7.1.1}$$

In the second type of simulation, we repeat the same kind of experiment, but assume the unbiased distribution of the covariates is unknown and estimate it with the algorithm proposed in the previous section. The third type of simulation explores the behaviour of $\mathcal{L}_I$ and $\mathcal{L}_J$ with a continuous covariate which follows a uniform density on the unit interval in the unbiased population, resulting in this mean lifetime bias density:

$$f_B(z;\alpha,\beta) = \frac{\beta}{\alpha}\frac{e^{-\frac{\beta z}{\alpha}}}{1-e^{-\frac{\beta}{\alpha}}}. \tag{7.1.2}$$

Once again, the unbiased density of the covariates is estimated.

For the first type of simulation, one thousand samples of size $n \in \{100, 1000\}$ were generated, for β values of $\{0, 1, 2, 4\}$, with Weibull parameters α and λ set to 2 and 1 respectively. The interest is to evaluate relative efficiency of $\mathcal{L}_J$ and $\mathcal{L}_I$ in estimating the covariate effect β in terms of ratio of mean squared error of the MLEs. A Newton-Raphson type of algorithm was used to maximize the likelihoods. The results are summarized in Table 7.1.

n	Censoring	β	Average $\hat{\beta}_J$	Average $\hat{\beta}_I$	S.d($\hat{\beta}_J$)	S.d.($\hat{\beta}_I$)	Eff.($\hat{\beta}_J : \hat{\beta}_I$)
100	$\sim$ 15%	0	0.011	0.003	0.165	0.179	1.172
	$\sim$ 30%	0	-0.005	-0.008	0.177	0.195	1.211
	$\sim$ 15%	1	1.030	1.025	0.193	0.200	1.067
	$\sim$ 30%	1	1.035	1.027	0.197	0.210	1.112
	$\sim$ 20%	2	2.066	2.055	0.276	0.284	1.039
	$\sim$ 15%	4	4.139	4.119	0.475	0.482	1.009
	$\sim$ 30%	4	4.165	4.149	0.509	0.511	0.988
1000	$\sim$ 15%	0	-0.001	-0.002	0.049	0.053	1.139
	$\sim$ 30%	0	-0.0002	-0.001	0.054	0.058	1.190
	$\sim$ 15%	1	1.006	1.006	0.060	0.063	1.107
	$\sim$ 30%	1	1.002	1.000	0.064	0.068	1.099
	$\sim$ 20%	2	2.009	2.006	0.083	0.087	1.084
	$\sim$ 15%	4	4.013	4.011	0.142	0.145	1.040
	$\sim$ 30%	4	4.013	4.011	0.151	0.154	1.044

Table 7.1: Efficiency results for discrete covariate with known $f_Z(z)$

A few comments are in order. For a discrete covariate with a known unbiased distribution, the MLEs obtained from $\mathcal{L}_J$ seem generally more efficient than those obtained using $\mathcal{L}_I$, though the method used seem to induce a slight bias (an overestimate), possibly because the longer living subpopulation tends to be more censored. The gain in efficiency seems also inversely proportional to $|\beta|$. This is expected due to the fact that the sampling proportion of the shorter lived subpopulation will go to zero as $|\beta| \to \infty$. Incidentally, it is not clear whether censoring has an effect on the relative efficiency of the estimates. While there could some bias caused by lopsided censoring as $|\beta| \longrightarrow \infty$, the variance of $\hat{\boldsymbol{\beta}}_J$ appears less affected by increases in censoring, as the information held in the sampling distribution of the covariates is not lessened by censoring. Note that the variance of the estimates increases with the amount of censoring and β.

For the second set of simulations, we kept the amount of censoring approximately to the same proportion ($\sim 30\%$), and reduced the number of simulations as estimating the unbiased distribution of the covariates is computationally time consuming. We also reduced the number of simulations per set of parameters to 100, with 1000 observations per simulation. The results are shown in Table 7.2.

β	Average $\hat{\beta}_J$	Average $\hat{\beta}_I$	S.D.$(\hat{\beta}_J)$	S.D.$(\hat{\beta}_I)$	Eff.$(\hat{\beta}_J : \hat{\beta}_I)$
0	0.006	0.004	0.055	0.052	0.915
1	0.997	0.995	0.066	0.064	0.958
2	2.006	2.008	0.084	0.083	0.987
4	3.996	4.002	0.150	0.149	0.992

Table 7.2: Efficiency results for discrete covariates with $f_Z(z)$

In the third type of simulations, 1000 simulations with 1000 observations per set of parameters were performed. The amount of censoring was kept around 30%. The results are displayed in Table 7.3.

β	Average $\hat{\beta}_J$	Average $\hat{\beta}_I$	S.d.$(\hat{\beta}_J)$	S.d.$(\hat{\beta}_I)$	Eff.$(\hat{\beta}_J : \hat{\beta}_I)$
0	-0.001	-0.001	0.102	0.102	1.000
1	1.001	1.001	0.108	0.108	1.000
2	2.014	2.014	0.115	0.115	1.000
4	3.996	3.996	0.162	0.162	1.000

Table 7.3: Efficiency results for continuous covariate

It is clear that the gain in efficiency we saw when we made the correct assumption about the unbiased distribution of the covariates is lost when we have to estimate it nonparametrically using the modification to Vardi's algorithm. Having to estimate $f_Z(z)$ induces variability in the estimation of the parameters. In the discrete case, 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. Note that the efficiency appears to increase with β. The greater the covariate effect, the more pronounced demarcation between the two subpopulation is, consequently, the estimated unbiased marginal distribution will be less affected by the inverse length-bias transformation. It should be mentioned that the estimates from the correct likelihood take much more computational resources.

For the continuous covariate case, the MLEs are practically identical across the scale (there were small numerical discrepancies that the table obfuscates through truncation of the displayed numbers), but the conditional likelihood is computationally much simpler and quicker. Also, in about 2% of the simulations, the Newton-Raphson method failed to iterate for the joint likelihood with continuous covariates, and therefore in those cases both sets of estimates were omitted from the table.

Though we used what appeared to be the most natural approach to implement the correct likelihood when the unbiased distribution of the covariates is unknown and did not show a gain in efficiency over the conditional approach, it appears feasible to gain efficiency using the correct likelihood provided we can develop a more efficient estimate of $f_{\mathbf{Z}}(\mathbf{z})$. Possible options include kernel density estimates or a Bayesian approach. The issue is how the parameters and unbiased marginal distribution of the covariates affect maximization the joint likelihood. When the covariate effects are small, the sampling marginal distribution of the covariates holds little information about the parameters, so that gain in information is small compared to the difficulties of working in a higher dimension. When the covariate effects are large, obtaining a representative sample becomes difficult as one subpopulation will be heavily underrepresented, reducing the effective sample size greatly.

7.2 Application to survival with dementia

In 1991 a nation wide prevalence study of dementia was carried out as part of the Canadian Study of Health and Aging. The first phase of the study was termed CSHA-1. Over a short period of time subjects were examined and classified into several dementia categories, including "probable Alzheimer's Disease", "possible Alzheimer's Disease", and "vascular dementia". All subjects with dementia were then followed until 1996 when the second phase began with a follow-up exam on those who were still alive, termed CSHA-2. Amongst the many aims of the study was the estimation of the survival distribution of individuals with dementia of any type, as well as an assessment of covariates that could possibly affect survival.

The relevant data for this aspect of the study included approximate date of onset, date of death or censoring and the usual death indicator. Subjects were censored if still alive at the end of the study or were lost to follow-up. There were 816 (possibly censored) observed survival times. The difficulty with such data is that they are length-biased, the diagnosis of dementia having been made on prevalent rather than incident cases. Since it can be reasonably assumed that the incident rate of dementia has remained constant, it follows that the times from onset of dementia to CSHA-1 (the random left truncation times) are uniformly distributed. That is, we may invoke the stationarity assumption referred to in the introduction. See Asgharian et al. (2006) for detailed discussion on stationarity.

Asgharian et al. (2002) proposed an unconditional approach to analyze right-censored length-biased data. They found the NPMLE of the survivor function and established its asymptotic behaviour. They compared their unconditional NPMLE with the conditional NPMLE (Wang, 1991), in the estimation of the survival function underlying the observed survival times of the CSHA. Wolfson et al. (2001), more concerned with substantive issues of dementia, analyzed the CSHA survival data using the *conditional* parametric MLE and the *conditional* NPMLE. They also assessed the effect of covariates on survival which were included in a Weibull model.

Here, we find the unconditional parametric (Weibull) MLE of the survival function of the group of patients diagnosed with dementia, in the CSHA using both the correct joint likelihood and the likelihood that is conditional on the covariates. It is easy to check that Assumptions A1-A6 are satisfied for the Weibull model. The details are omitted. Finally we assess the effect of type of dementia on survival from onset, by estimating the parameters of a length-biased Weibull model with the covariate, "final diagnosis".

Let the covariates Z_r, $r = 1, 2, 3$, be defined as follows:

$$Z_r = \begin{cases} 1, & \text{if the subject has the rth disease} \\ 0, & \text{if the subject does not have the rth disease} \end{cases}$$

where 1="probable Alzheimer's", 2="possible Alzheimer's", and 3="vascular dementia". In our application (Z_1, Z_2, Z_3) reduce to a vector of dimension 2, as the covariates for "probable Alzheimer's" gets absorbed into the baseline hazard. For the ith subject:

$$\mathbf{Z}_i = \begin{cases} (0,0), & \text{if the subject has ``probable Alzheimer's''} \\ (1,0), & \text{if the subject has ``possible Alzheimer's''} \\ (0,1), & \text{if the subject has ``vascular dementia''} \end{cases}$$

Asymptotic confidence intervals were found using the observed information matrix, which is justified by the results given in section 4 and 5. For the joint likelihood, we estimated $f_{\mathbf{Z}}(\mathbf{z})$ using the method proposed in section 7. The time scale for survival was set in years. The results are given in Table 7.4 and Table 7.5.

Parameter	Estimate	Standard Deviation	Confidence Interval
α	1.226	0.046	(1.136, 1.317)
λ	0.208	0.010	(0.188, 0.228)
β_2	-0.150	0.065	(-0.277, -0.024)
β_3	0.073	0.072	(-0.069, 0.214)
$\exp(\beta_2)$	0.861		(0.758, 0.977)
$\exp(\beta_3)$	1.075		(0.934, 1.238)

Table 7.4: Parameter estimates and 95% confidence intervals using $\mathcal{L}_I$

Parameter	Estimate	Standard Deviation	Confidence Interval
α	1.226	0.002	(1.223, 1.229)
λ	0.208	0.0004	(0.207, 0.209)
β_2	-0.150	0.002	(-0.155, -0.146)
β_3	0.073	0.003	(0.068, 0.078)
$\exp(\beta_2)$	0.861		(0.857, 0.864)
$\exp(\beta_3)$	1.075		(1.070 , 1.081)

Table 7.5: Parameter estimates and 95% confidence intervals using $\mathcal{L}_J$

Comparing the two tables, we see that the estimates from either likelihood are practically identical. The correct likelihood however appears to be uncannily precise, but that is probably a numerical optimization issue.

We checked the stationarity assumption using the method suggested by Wang(1991), to estimate the distribution function of the left truncation times, and found this assumption to be reasonable (see Asgharian et al., 2002). New methods devised by Asgharian, et al (2006) to check stationarity of incidence rate has further validated this assumption for the CSHA data.

Figure 7.1 below compares the unbiased nonparametric estimate of the survival curve to the parametric estimate.

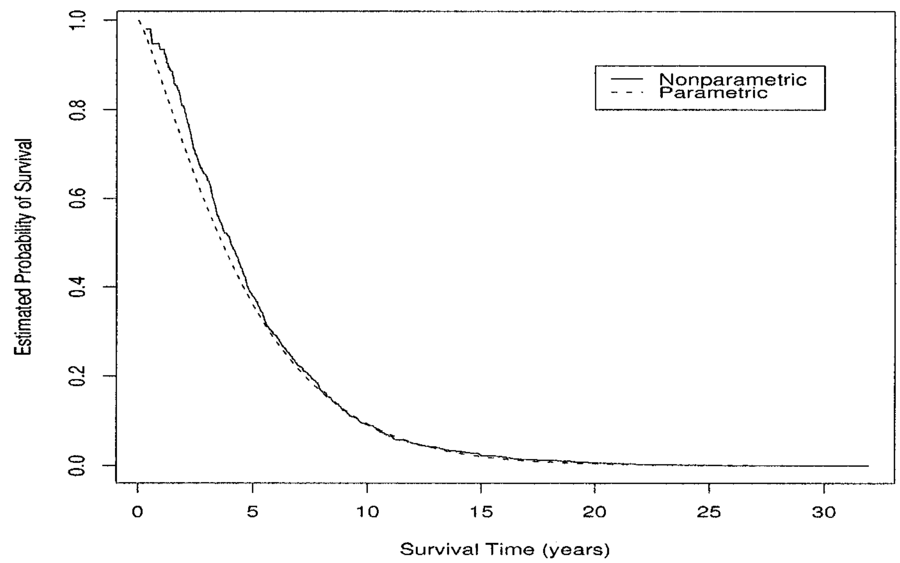

Figure 7.1: Nonparametric and parametric estimates of S(x)

Eschewing Andrews and Buchinsky's aforementioned guidelines for the number of bootstrap repetitions (as their methods only cover standard bootstrap techniques, and hence not this semiparametric approach to length-biased right censored Weibull data), we opted to do one thousand bootstrap samples with 816 observations (the original sample size), and estimated the model parameters using $\mathcal{L}_I$ and $\mathcal{L}_J$ in each case. There is one caveat: the unbiased distribution of the covariates estimated using Algorithm 1 from the original data was assumed to be the true unbiased distribution for all bootstrap samples (hence it was not re-estimated in the maximization of $\mathcal{L}_J$).

The bootstrap parameter estimates are given in Tables 7.6 and 7.7 be-

Parameter	Value	Estimate	Standard Deviation	Bias
α	1.226	1.234	0.050	0.008
λ	0.208	0.207	0.011	-0.001
β_2	-0.150	-0.151	0.067	-0.001
β_3	0.073	0.075	0.072	0.003

Table 7.6: Bootstrap parameter estimates from $\mathcal{L}_I$

Parameter	Value	Estimate	Standard Deviation	Bias
α	1.226	1.234	0.050	0.008
λ	0.208	0.206	0.010	-0.002
β_2	-0.150	-0.132	0.047	0.019
β_3	0.073	0.094	0.050	0.021

Table 7.7: Bootstrap parameter estimates from $\mathcal{L}_J$

low. The relative efficiency of the bootstrap parameter estimates is given in Table 7.8 in terms of ratio of variances and in terms of ratio of mean squared errors (as the bootstrap estimates appear to be biased). For each set of parameters, a survival curve from the mixture of Weibull was computed (using the unbiased proportion of covariates estimated from the original data set), the point-wise confidence bands displayed in Figure 7.2 below were obtained by taking the desired quantiles of the bootstrap survival curves at each sample time point. Finally, though the confidence bands for the survival curves obtained through $\mathcal{L}_J$ and $\mathcal{L}_I$ appear indistinguishable, Figure 7.3 gives the relative efficiency of the joint approach over the conditional approach for the survival function at each sample time point.

Parameter	α	λ	β_2	β_3
$\text{Var}(\boldsymbol{\theta}_{I,i})/\text{Var}(\boldsymbol{\theta}_{J,i})$	1.000	1.140	2.032	2.015
$\text{MSE}(\boldsymbol{\theta}_{I,i})/\text{MSE}(\boldsymbol{\theta}_{J,i})$	1.002	1.086	1.754	1.711

Table 7.8: Relative efficiency of $\mathcal{L}_J$ compared to $\mathcal{L}_I$ for the bootstrap parameter estimates

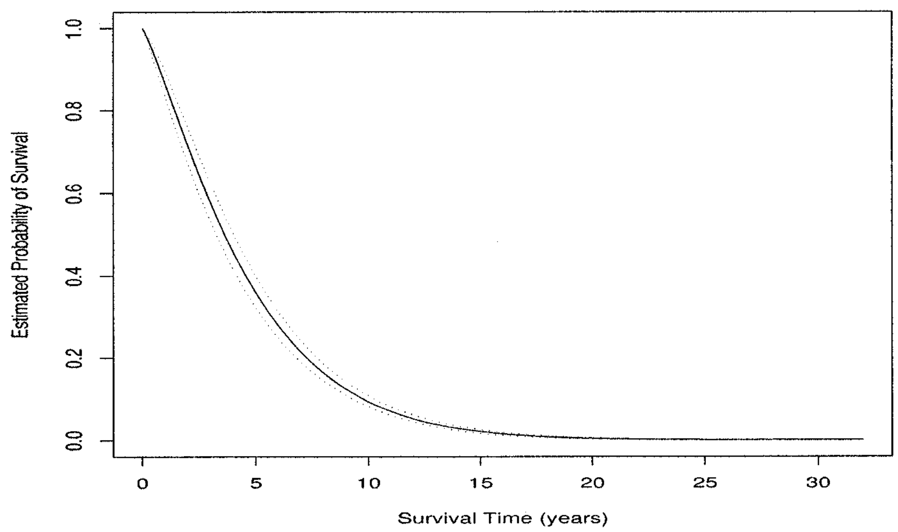

Figure 7.2: Estimated parametric survival function and its bootstrap pointwise confidence bands

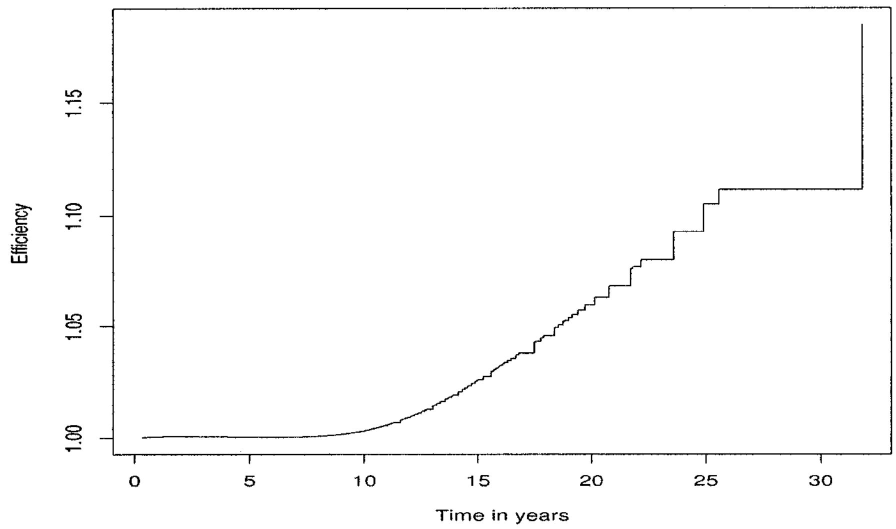

Figure 7.3: Estimated efficiency of $\mathcal{L}_J$ over $\mathcal{L}_I$ for estimation of the survival function over time

A few comments are in order. The first thing to be noticed is that the bootstrap standard deviations for $\mathcal{L}_I$ match the estimated standard deviations from the original data, while those for $\mathcal{L}_J$ differ (the original estimates being dubious because of numerical issues related to the optimization criterion). While the correct likelihood gives less variable estimates than the incorrect one, it is more prone to bias when the unbiased distribution of the covariates is fixed in the optimization. The bias might be due to the mixture effect when the covariate effects are small. The shorter observed lifetimes carry more weight in the length-bias to unbiased conversion, both for the lifetime distribution and the unbiased covariate distribution. When the subpopulations are distinct (i.e. the covariate effects are away from zero), only one subpopulation will consistently have shorter observed lifetime than the others and the unbiased covariate distribution will appear the stable over many samples. When the covariate effects are small, the estimated unbiased distribution of the covariates will be skewed toward whichever subpopulation had shorter lifetimes in a sample, and so the unbiased covariate distribution will be harder to estimate consistently.

To illustrate this, we use two simulated samples with only one covariate, one with a very small effect ($\beta = 0.1$) and one with a large effect ($\beta = 4$). We estimate the parameters and the unbiased distribution of the covariate then generate one thousand new samples through our semiparametric bootstrap method and obtain new bootstrap parameter estimates to look at the bias for $\hat{\beta}_I$ and $\hat{\beta}_J$ compared to the β used to generate the bootstrap samples (not the true β of the original population). The results are given in Table 7.9.

β	Bias($\hat{\beta}_I$)	Bias($\hat{\beta}_J$)
0.1	0.003	-0.019
4	-0.013	0.015

Table 7.9: Comparison of bias in bootstrap estimates

Note that for small β, $\hat{\beta}_I$ has almost no bias, which is expected as $\mu(Z = 1; \boldsymbol{\theta}) \approx \mu(Z = 0; \boldsymbol{\theta}) \approx \mu(\boldsymbol{\theta})$ and $\mathcal{L}_I$ is almost equal to $\mathcal{L}_J$ (they would be equal if β was exactly 0). The bias in $\hat{\beta}_J$ is significant for small covariate effect. On the other hand, for a large effect, the biases are comparable, though the absolute bias has increased for $\hat{\beta}_I$ but decreased for $\hat{\beta}_J$.

The efficiency curve is rather interesting. It appears that the joint likelihood approach gives indistinguishable results compared to the conditional approach for the first 9 years, then gradually becomes increasingly more efficient after this point, making it overall more efficient, as suggested by looking at the bootstrap parameters estimates.

Chapter 8

What ended up on the cutting room floor

In this chapter we overview what avenues of research were explored in the development of this thesis but did not yield much fruitful results. Though they were abandoned for various reasons, they might still prove useful in the future.

8.1 A semiparametric approach

This thesis was originally meant to expand the work done in Bergeron (2003), and find a semiparametric approach for regression in the context of length-biased and right-censored lifetime data analysis. This could also be taken as an extension of the semiparametric approach of Cox (1969), as it is based on proportional hazard models and profile likelihood. The main difficulty of this approach is that, when one does not impose a parametric model on the baseline hazard, the conditional mean lifetime (given the covariates) does not have a closed form. Writing the likelihood in terms of hazard:

$$\begin{aligned}\mathcal{L}_J(\mathbf{x}, \mathbf{Z}; \boldsymbol{\beta}) &= \prod_{i=1}^{n} \frac{\left(e^{\boldsymbol{\beta}'\mathbf{z}_i}\lambda_0(x_i)\right)^{\delta_i} \exp\left(-e^{\boldsymbol{\beta}'\mathbf{z}_i}\Lambda_0(x_i)\right)}{\mu(\boldsymbol{\beta})} \\ &= \prod_{i=1}^{n} \frac{\left(e^{\boldsymbol{\beta}'\mathbf{z}_i}\lambda_0(x_i)\right)^{\delta_i} \exp\left(-e^{\boldsymbol{\beta}'\mathbf{z}_i}\Lambda_0(x_i)\right)}{\displaystyle\int_{\mathbf{z}} \mu(\mathbf{z}; \boldsymbol{\beta}) f_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z}},\end{aligned} \tag{8.1.1}$$

where

$$\mu(\mathbf{z}; \boldsymbol{\beta}) = \int_0^{\infty} \exp\left(-e^{\boldsymbol{\beta}'\mathbf{z}}\Lambda_0(x)\right) dx. \tag{8.1.2}$$

Minimally, this requires $\mu(\mathbf{z}_i; \boldsymbol{\beta})$ for every $\mathbf{z}_i$ in the sample.

One possible approach to find this mean is the use of an approximation. This would result in an approximate likelihood, such as discussed by Chen and Jennrich (2002). The ostensible choice is a Laplace type of approximation, which applies to integrals of the form:

$$\int_D \exp\left(-ng(x)\right) dx, \tag{8.1.3}$$

where D is some domain in R^p and $g(x)$ is a unimodal infinitely differentiable real valued function with a unique minimum at some point x_0, and n is some large constant (in statistical contexts, it is often the sample size). In

8.1.3, $g(x)$ is the cumulative hazard, which is monotone increasing and has a minimum at $x = 0$ (this is a one-dimensional integral over time). However, in order to use a Laplace approximation, some constraints have to be imposed on $\Lambda_0(x)$.

One such approximation is given by Erdélyi (1956, pp. 36-38), requires $\Lambda_0(x)$ to be approximately polynomial at the origin. This approximation actually reduces to imposing a model which behaves like a Weibull model close to the origin.

Another possibility is to consider the conditional likelihood $\mathcal{L}_I$, if one could assume that the results of Theorem 1 hold for an infinite dimensional parameter space. While it should not yield the most efficient estimates, its formulation might be more suitable for algebraic manipulation, as it works in a space of smaller dimension than $\mathcal{L}_j$:

$$\begin{aligned}\mathcal{L}_I(\mathbf{x}, \mathbf{Z}; \boldsymbol{\beta}) &= \prod_{i=1}^{n} \frac{\left(e^{\boldsymbol{\beta}'\mathbf{z}_i}\lambda_0(x_i)\right)^{\delta_i} \exp\left(-e^{\boldsymbol{\beta}'\mathbf{z}_i}\Lambda_0(x_i)\right)}{\mu(\mathbf{z}_i; \boldsymbol{\beta})} \\ &= \frac{\prod_{i=1}^{n} \left(e^{\boldsymbol{\beta}'\mathbf{z}_i}\lambda_0(x_i)\right)^{\delta_i} \exp\left(-e^{\boldsymbol{\beta}'\mathbf{z}_i}\Lambda_0(x_i)\right)}{\prod_{i=1}^{n} \int_0^{\infty} \exp\left(-e^{\boldsymbol{\beta}'\mathbf{z}}\Lambda_0(x_i)\right) dx_i}.\end{aligned} \tag{8.1.4}$$

using Fubini's theorem (see Billingsley 1995, p. 233) the denominator can be rewritten as

$$\prod_{i=1}^{n} \mu(\mathbf{z}_i; \boldsymbol{\beta}) = \int_0^\infty \dots \int_0^\infty \exp\Big(- \sum_{i=1}^{n} e^{\boldsymbol{\beta}'\mathbf{z}_i} \Lambda_0(x_i) \Big)\, dx_1 \dots dx_n, \qquad (8.1.5)$$

which is an integral in R^n. This allows to use another Laplace approximation given by Wong (1989, pp. 494-498), under some extra conditions on Λ_0, namely that $0 < \lambda_0'(0) < +\infty$. Extra care should be taken when such an assumption is imposed as it effectively eliminates the possibility of a Cox model (where the estimated hazard is null except at event times). It also eliminates any Weibull model with shape parameter $\alpha > 2$ (in fact, this approximation by Wong is equivalent to a Weibull model with $\alpha = 2$). Furthermore, as the dimension of the integral increase linearly with n, there is no hope for it to converge with an increasing sample size, as Shun and McCullough (1995) have shown that a multidimensional Laplace approximation can converge only if the number of integrals is $\mathrm{O}(n^{1/2})$. It might be worthwhile to look at a generalized version of Laplace's method, such as proposed by Fu and Wong (1980) or the use of asymptotic modes and fully exponential Laplace approximations introduced by Miyata (2004).

Since the approximation approach did not prove fruitful, obtaining $\mu(\mathbf{z}_i; \boldsymbol{\beta})$ through computational methods was considered. The idea was to use jackknife pseudo-values such as proposed by Andersen, Klein and Rosthøj (2003). While an elegant approach to dealing with $\mu(\boldsymbol{\theta})$ and $\mu(\mathbf{z}; \boldsymbol{\theta})$, it cannot be easily adjusted in the presence of length bias. The issue is that, though one can nonparametrically obtain $\hat{\mu}(\boldsymbol{\beta})$ and a leave-one-out $\hat{\mu}_{-i}(\boldsymbol{\beta})$, using

$$\hat{\mu}(\mathbf{z}_i; \boldsymbol{\beta}) = n\hat{\mu}(\boldsymbol{\beta}) - (n-1)\hat{\mu}_{-i}(\boldsymbol{\beta}) \tag{8.1.6}$$

will not yield a conditional mean that is necessarily positive. Various similar tricks, such as doing jackknife in the length-biased space, *then* using inverse length bias transformation, grouping observations with same $\mathbf{z}$, working with survival functions, or using some basic transformations do not result in estimates exhibiting the necessary properties. The two main causes of these setbacks are that we are dealing with bounded quantities and that each observation does not carry the same weight in the unbiased population, together making any use of jackknife a delicate endeavour. We hope to explore this line of thought further in the future.

As a semiparametric approach involves an infinite dimensional hazard or a profile likelihood, there are still many directions to explore. Goldwasser, Tian and Wei (2004) proposed methods of estimating cumulative hazard through asymptotically pivotal estimating functions. Barndorff-Nielsen and Jupp (1988) used a geometric approach for models involving profile likelihood, while Severini and Wong (1992) developed a general approach to estimate the parametric components in semiparametric models. Fan, Gijbels and King (1997) estimated hazard locally in both parametric and semiparametric proportional hazard models. If the conditional approach can be considered a valid option, Cox and Reid (1987) offered an approximate method of conditional inference through parameter orthogonality. Another option is to investigate residuals in length-biased data, thereby extending the work of Schoenfeld (1982).

8.2 A measure of dependence for length-biased right-censored data

Without a semiparametric model, the adaptation of the information gain based measure of dependence of Kent (1983) to uniformly left-truncated and right-censored data has to be reduced to parametric models. Kent and O'Quigley (1988) used Fraser information (Fraser 1965) to extend the work of Linfoot (1957) (itself based on Shannon information, see Khinchin 1957) and introduced methods that could take into account censoring in both parametric and semiparametric models to obtain a measure of dependence for survival data. This could be accomplished as estimating information gain reduces to a problem of maximum likelihood, or of partial likelihood in the case of Cox's model. However, this particular measure is based on a conditional model hence it cannot apply for length-biased data through a trivial modification.

Instead, we go back to Kent's ρ_J^2 measure based on a joint model for the variable of interest X and the covariates $\mathbf{Z}$. Consider Y to be a random variable with true distribution $g(y)dy$. Suppose we have two families of parametric models $\{f(y;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_i\}$, $i \in \{0,1\}$, and $\Theta_0 \subset \Theta_1$. Let $\boldsymbol{\theta}_i$ be the maximizer of the expected log-likelihood,

$$\ell(\boldsymbol{\theta}) = \int \log f(y;\boldsymbol{\theta})g(y)\,dy, \tag{8.2.1}$$

for $\boldsymbol{\theta} \in \Theta_i$. Note that 8.2.1 is called the Fraser information and $\boldsymbol{\theta}_i$ is a theoretical equivalent to a maximum likelihood estimate. Define the information gain to be twice the Kullback-Leibler (1951) information gain (γ will be used instead of Γ to avoid confusion with Euler's Gamma function):

$$\gamma(\boldsymbol{\theta}_1 : \boldsymbol{\theta}_0) = 2\big[\ell(\boldsymbol{\theta}_1) - \ell(\boldsymbol{\theta}_0)\big]. \tag{8.2.2}$$

Note that $\gamma(\boldsymbol{\theta}_1 : \boldsymbol{\theta}_0)$ increases as the family $\{f(y;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_0\}$ approaches $g(y)dy$ compared to $\{f(y;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_0\}$, and is analogous to $2\log\Lambda$ where Λ is the generalized likelihood ratio. Now consider $Y = (X, \mathbf{Z})$ and Θ_0 to be the parameter space in which X and $\mathbf{Z}$ are independent, then $\gamma(\boldsymbol{\theta}_1 : \boldsymbol{\theta}_0)$ gives the information gain from an independent model to one with dependence. Kent's joint correlation coefficient is then

$$\rho_J^2(X, \mathbf{Z}) = 1 - \exp\big(-\gamma(-\boldsymbol{\theta}_1 : \boldsymbol{\theta}_0)\big). \tag{8.2.3}$$

When we have length-biased lifetime data with right censoring, one can obtain an estimate $\hat{\boldsymbol{\theta}}_1$ through maximization of $\mathcal{L}_J$, while Θ_0 is equivalent to setting the covariate effects $\boldsymbol{\beta}$ to zero. Kent suggests using a "fitted" estimate for $\boldsymbol{\theta}_0$, using the fitted lifetime values to accommodate censored observations:

$$\tilde{\boldsymbol{\theta}}_0 = \arg\max_{\boldsymbol{\theta}\in\Theta_0} \frac{1}{n}\sum_{i=1}^{n} \int \log\left(f_{LB}(x, \mathbf{z}_i; \boldsymbol{\theta})\right) f_{LB}(x, \mathbf{z}_i; \hat{\boldsymbol{\theta}}_1)\, dx. \tag{8.2.4}$$

Similarly, there is a "fitted" estimate for information gain:

$$\tilde{\gamma} = \frac{2}{n}\sum_{i=1}^{n} \int \log\left(f_{LB}(x, \mathbf{z}_i; \boldsymbol{\theta}_1)/f_{LB}(x, \mathbf{z}_i; \tilde{\boldsymbol{\theta}}_0)\right) f_{LB}(x, \mathbf{z}_i; \hat{\boldsymbol{\theta}}_1)\, dx. \tag{8.2.5}$$

Implementing this measure of dependence and ascertaining its asymptotic properties will be pursued in the near future.

Chapter 9

New horizons

To bring this thesis to a closing point, we reflect on what has been accomplished and explore in which directions this research will lead us.

In the beginning, we started with a regression problem: how to measure the effect of covariates on the lifetime of subjects sampled from prevalent cohorts subject to right censoring. Under the stationarity assumption of the onset times, the observed lifetimes are length-biased, which we showed implies that the covariates are also sampled with a bias and are informative of the parameters that are to be estimated. While reviewing the literature, it appeared that this natural feature of biased sampling had been largely ignored in the context of survival analysis. Expanding our focus to include more general biased sampling issues allowed us to see more clearly how to construct the proper likelihood, going from a traditional conditional approach to a joint modeling point of view. By doing so, the insidious nature of making the mistake of ignoring the sampling distribution of the covariates was exposed: this misguided approach still yields asymptotically valid point estimates. And so it was demonstrated that the correct approach gives more efficient parameter estimates. We derived the asymptotic properties of the MLE and adapted widely used models for incident population to length-biased populations through this joint likelihood approach. In the interest of implementation, a number of algorithms were introduced: a modification of Vardi's EM algorithm to recover the unbiased distribution of the covariates nonparametrically, algorithms to generate length-biased lifetimes with covariates and a semiparametric bootstrap algorithm to get confidence bands for the estimated survival function. We analyzed the performance of these through simulation and applied our theory to the CSHA data. Finally, we discussed what appeared to be dead ends at the moment in our research.

But what has been accomplished so far is only to scratch the surface of this investigation. There are numerous avenues of research from this point. Certainly, a number of theoretical results regarding rates of convergence of estimators are already being worked on. But what appears to be most needed is a systematic approach to implement the likelihood. In the light of the relationship between length-bias in the lifetime and mean lifetime bias in the covariates, the relationship between the lifetime distribution and the covariate distribution should be explored further, to see what can be generalized and to find formulae for $f_B(\mathbf{z};\boldsymbol{\theta})$ given different $f_{\mathbf{Z}}(\mathbf{z})$. To that effect, in parametric models, a Bayesian approach could be useful. Another approach could be more computational, such as the use of kernel density estimators. Still, going from a likelihood to an estimating equations approach seems a fruitful possibility which will be the next task after this thesis.

There is also the issue of models. So far, proportional hazard models and accelerated failure time models were considered. However, there are many other possibilities to extend the applicability of our approach. Some possible developments include accelerated hazard models, proposed by Chen and Wang (2000) and discussed by Chen and Jewell (2001). These models can be useful in clinical trials where treatment necessitates some time to be fully effective. Other models in prevalent cohorts focus on residual lifetime, such as the proportional mean residual lifetime models (see Oakes and Dasu 2003; Chen and Cheng 2005; Chen et al. 2005). As mean lifetimes are sensitive to length-bias, it could be possible to develop a median residual lifetime approach that should theoretically be more robust. As mentioned in the previous chapter, a satisfactory extension of Cox's semiparametric proportional hazard model remains the Holy Grail of the subject.

On the practical side, the computational aspects of the problems could be investigated further. There are difficulties in finding a general method to write down the likelihood so that to minimize numerical problems. Subsequently, this might require the creation of new optimization algorithms specifically designed for this setting, as the currently available methods are not completely up to the task. A Monte-Carlo approach, such as suggested by Geyer (1994) could be considered.

One could always try to relax the assumption of stationarity and see how this affects the behaviour of the bias both in the lifetime and the covariates. This would complexify the likelihood in this setting. Another relaxable assumption of our problem is to have a mixture of unbiased and biased observations. Asymptotics for the likelihood in this case have not been assessed.

Finally, while the tools are being developed, they also need to be used in real applications. The combination of length-bias and covariates is not limited to lifetime data analysis and the methods developed in this thesis could extend to other topics. These include genetics (see Terwilliger et al. 1997) and problems in economics such as unemployment data. Further applications involve longitudinal data. As suggested by McFadden (1962) there are cases of recurrent events where the initial observation may be biased while subsequent events are not, and there exists problems of biased follow-up, as in Lin, Scharfstein and Rosenheck (2004).

Bibliography

[1] Addona, V. (2001). *Group Comparison in the Presence of Length-Biased Data.* Unpublished MSc thesis, *McGill University Department of Mathematics and Statistics.*

[2] Addona, V. and Wolfson, D. B. (2006). "A formal test for the stationarity of the incidence rate using data from a prevalent cohort study with follow-up." *Lifetime Data Analysis.*

[3] Alioum, A. and Commenges, D. (1996). "A proportional hazards model for arbitrarily censored and truncated data." *Biometrics*, 52, 512-524.

[4] Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical models based on counting processes.* Springer Series in Statistics. Springer-Verlag.

[5] Andersen, P. K., Klein, J. P. and Rosthøj, S. (2003). "Generalised linear models for correlated pseudo-observations, with applications to multi-state models." *Biometrika*, 90, 15-27.

[6] Andrews, D. W. K. and Buchinsky, M. (2000). A Three-step method for choosing the number of bootstrap repetitions. *Econometrica* 68, 23-51.

[7] Asgharian, M.(2001). "On the Singularities of the Information Matrix with an Application to Mixture Distributions." Technical Report No. 2001-03, *McGill University Department of Mathematics and Statistics.*

[8] Asgharian, M. and Wolfson, D. B. (2000). "Modeling Covariance in Multi-path Changepoint Problems (II) Consistency of the MLE." Technical Report No 2000-04. *Department of Mathematics and Statistics, York University.*

[9] Asgharian, M. and Wolfson, D. B. (2001a). "Asymptotic behaviour of the $NPMLE$ of the survivor function when the data are length-biased and subject to right censoring." , Technical Report No. 2001-01, *McGill University Department of Mathematicsand Statistics*

[10] Asgharian, M. and Wolfson, D. B. (2001b). "Covariates in multipath change-point problems: modelling and consistency of the MLE." *The Canadian Journal of Statistics*, 29, 515-528.

[11] Asgharian, M., Wolfson, D. B. and M'Lan, C. E. (2002). "Inference Based on Cross-Sectional Sampling for Diseases with Stationary Incidence." Technical Report No. 2002-01, *McGill University Department of Mathematicsand Statistics*

[12] Asgharian, M., M'Lan, C. E. and Wolfson, D. B. (2002). "Length-biased sampling with right censoring: An unconditional approach." *Journal of the American Statistical Association*, 97, 201-209.

[13] 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." *Annals of Statistics*, 33, 2109-2131.

[14] Asgharian, M., Wolfson, D. B. and Zhang, X. (2006). "Checking stationarity of the incidence rate using prevalent cohort data." *Statistics in Medicine*, 25, 1751-1767.

[15] Baker, S. G., Fitzmaurice, G. M., Freedman, L. S. and Kramer, B. S. (2006). "Simple adjustments for randomized trials with nonrandomly missing or censored outcomes arising from informative covariates." *Biostatistics* 7, 29-40.

[16] Barndorff-Nielsen, O. E., and Jupp, P. E. (1988). "Differential Geometry, Profile Likelihood, L-Sufficiency and Composite Transformation Models." *The Annals of Statistics*, 16, 1009-1043.

[17] Begg, C. B. and Gray, R. J. (1987). "Methodology for case-control studieswith prevalent cases." *Biometrika*, 74, 191-195.

[18] Begg, C. B.and Greenes, R. A. (1983). "Assesment of Diagnostic Tests When Disease Verification is Subject to Selection Bias." *Biometrics*, 39, 207-215.

[19] Bender, R. , Augustin, T. and Blettner, M. (2005). "Generating survival times to simulate Cox proportional hazards models." *Statistics in Medicine*, 24, 1713-1723.

[20] Berger, A., Bodian, C. A., and Hirsch, W. M. (1996). "On Estimating Incidence Rates of Diseases With Delayed Onset Using Biased Samples". *Journal of the American Statistical Associaton*, 91, 831-840.

[21] 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.*

[22] Bickel, P. J. and Ritov, J. (1991). "Large Sample Theory of Estimation in Biased Sampling Regression Models, I." *The Annals of Statistics*, 19, 797-816.

[23] Bilker, W. B. and Wang, M.-C. (1997). "Bootstrapping left truncated and right censored data". *Communications in Statistics. Simulation and Computation* 26, 141-171.

[24] Billingsley, Patrick (1995). *Probability and Measure.* Third edition. John Wiley & Sons, New York.

[25] Blumenthal, S. (1967). "Proportional sampling in life length studies." *Technometrics*, 9, 205-218.

[26] Breslow, N. E. (2003). "Are Statistical Contributions to Medicine Undervalued?" *Biometrics*, 59, 1-8.

[27] Breslow, N. and Holubkov, R. (1997). "Maximum Likelihood Estimation of Logistic Regression Parameters under Two-phase, Outcome-dependent Sampling." *JRSS B*, 59, 447-461.

[28] Breslow, N., McNeney, B. and Wellner, J. A. (2003). "Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling." *Annals of Statistics*, 31, 1110-1139.

[29] Brookmeyer, R. and Gail, M. H. (1987). "Biases in Prevalent Cohorts." *Biometrics*, 43, 739-749.

[30] Chen, J.-S. and Jennrich, R. I. (2002). "Simple Accurate Approximations to Likelihood Profiles." *Journal of Computational and Graphical Statistics*, 11, 714-732.

[31] Chen, K. (2001). "Parametric models for response biased sampling." *JRSS B*, 63, 775-789.

[32] Chen, Y. Q. and Cheng, S. (2005). "Semiparametric regression analysis of mean residual with censored survival data." *Biometrika*, 92, 19-29.

[33] Chen, Y. Q. and Jewell, N. P. (2001). "On a general class of semiparametric hazards regression models." *Biometrika*, 88, 687-702

[34] Chen Y. Q., Jewell N. P., Lei X. and Cheng, S. C. (2005). "Semiparametric estimation of proportional mean residual life model in presence of censoring." *Biometrics*, 61, 170-178.

[35] Chen, Y. Q. and Wang, M.-C. (2000). "Analysis of Accelerated Hazards Models." *JASA* 95, 608-618.

[36] Cnaan, A. and Ryan, L.(1989). "Survival Analysis in Natural History Studies of Disease." *Statistics in Medicine*, 8, 1255-1268.

[37] Correa, J.A. and Wolfson D.B. (1999). "Length-Bias: Some Characterizations and Applications." *J. Statist. Comput. Simul.*, 64, 209-219.

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

[39] Cox, D. R. (1972). "Regression models and life tables (with discussion)." *J. R. Statist. Soc. B*, 34, 187-220.

[40] Cox, D. R. and Reid, N. (1987). "Parameter Orthogonality and Approximate Conditional Inference." *JRSS B*, 49, 1-39.

[41] Cristóbal, J. A. and Alcalá, J. T. (2000). "Nonparametric regression estimators for length biased data." *Journal of Statistical Planning and Inference*, 89, 145-168.

[42] Cristóbal, J. A., Ojeda, J. L. and Alcalá, J. T. (2004)."Confidence Bands in Nonparametric Regression With Length Biased Data", *Annals of the Institute of Statistical Mathematics*, 56, 475-496.

[43] Davidov, O. and Zelen, M. (2001). "Referent sampling, family history and relative risk: the role of length-biased sampling." *Biostatistics*, 2, 173-181.

[44] Efron, B. (1981). "Censored Data and the Bootstrap." *Journal of the American Statistical Association*, 76, 312-321.

[45] Efron, B. (1987). "Better Bootstrap Confidence Intervals (with discussion)." *Journal of the American Statistical Association*, 82, 171-200.

[46] Efron, B. and Tibshirani, R. (1986). "Bootstrap Methods for Standard Errors, Confidence Intervals and Other Measures of Statistical Accuracy." *Statistical Science*, 1, 54-77.

[47] Erdélyi, A. (1956). *Asymptotic Expansions.* Dover Publications, New York.

[48] Fan, J, Gijbels, I. and King, M. (1997). "Local likelihood and local partial likelihood in hazard regression." *Annals of Statistics*, 92, 1661-1690.

[49] Fisher, R. A. (1934). "The effects of methods of ascertainment upon the estimation of frequencies." *Annals of Eugenics*, 6, 13-25.

[50] Folland, G. B. (1999) *Real Analysis: Modern Techniques and Their Applications.* Wiley Interscience.

[51] Fraser, D. A. S. (1965). "On information in statistics." *Ann. Math, Statist.*, 36, 890-896.

[52] Fu, J. C. and Wong, R. (1980). "An Asymptotic Expansion of a Beta-Type Integral and its Application to Probabilities of Large Deviations." *Proceedings of the Amercian Mathematical Society*, 79, 410-414.

[53] Gao, S. and Hui, S. (2000). " Estimating the incidence of dementia from two-phase sampling with non-ignorable missing data." *Statistics in Medicine*, 19, 1545-1554.

[54] Geyer, C. J. (1994). "On the Convergence of Monte Carlo Maximum Likelihood Calculations." *JRSS B*, 56, 261-274.

[55] Gilbert, P. B., Lele, S. R. and Vardi, Y. (1999). "Maximum likelihood estimation in semiparametric selection bias models with application to an AIDS vaccine." *Biometrika*, 86, 27-43.

[56] Gill, R. D., Vardi, Y. and Wellner, J. A. (1988). "Large sample theory of empirical distributions in biased sampling models." *The Annals of Statistics*, 16, 1069-1112.

[57] Glesby, M. J., and Hoover, D. R. (1996). "Survivor treatment selection bias in observational studies: examples from the AIDS literature." *Annals of Internal Medicine*, 124, 999-1005.

[58] Goldsmith, P. L.(1967). "The calculation of true particle size distributions from the sizes observed in a thin slice." *British Journal of Applied Physics*, 18, 813-830.

[59] Goldwasser, M. A., Tian, L. and Wei, L. J. (2004). "Statistical inference for infinite-dimensional parameter via asymptotically pivotal estimating functions." *Biometrika*, 91, 81-94.

[60] Hall, P., Härdle, W., Kleinow, T. and Schmidt, P. (2000). "Semiparametric bootstrap approach to hypothesis tests and confidence intervals for the Hurst coefficient." *Statistical Inference for Stochastic Processes*, 3, 263-276.

[61] Huang, Y. and Wang, M.-C. (1995). "Estimating the Occurance Rate for Prevalent Survival Data in Competing Risks Models." *JASA*, 90, 1406-1415.

[62] Huang, Chiung-Yu and Wang, M.-C. (2004). "Joint Modeling and Estimation for Recurrent Event Processes and Failure Time Data." *JASA*, 99, 1153-1164.

[63] Ibragimov, I.A. and Has'minskii, R.Z.(1981). *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York.

[64] Kaplan, E. L. and Meier, P. (1958). "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association*, 53, 457-481.

[65] Kent, J. T. (1983). *Information Gain and a General Measure of Correlation.* Biometrika, Vol. 70, No. 1, pp 163-173.

[66] Kent, J. T. and O'Quigley, J. (1988). "Measures of Dependence for Censored Survival Data." *Biometrika*, 75, 525-534.

[67] Khinchin, A. I. (1957). *Mathematical Fundations of Information Theory.* Dover Publications, New York.

[68] Kullback, S. and Leibler, R. A. (1951). "On information and sufficiency." *Ann. Math. Statist*, 22, 79-86.

[69] Lagakos, S. W., Barraj, L. M. and De Gruttola, V. (1988). "Nonparametric analysis of truncated survival data, with applications to AIDS." *Biometrika*, 75, 515-523.

[70] Lawless, J.F., Kalbfleisch, J. D. and Wild, C. J. (1999). "Semiparametric methods for response-selective and missing data problems in regression." *JRSS B*, 61, 413-438.

[71] Lin, H., Scharfstein, D. O. and Rosenheck. R. A. (2004). "Analysis of longitudinal data with irregular, outcome-dependent follow-up." *JRSS B*, 66, 791-813.

[72] Linfoot, E. H. (1957). "An Informational Measure of Correlation." *Information and Control*, pp 85-89.

[73] Martin, E. C., Betensky, R. A. (2005). "Testing quasi-independence of failure and truncation times via conditional Kendall's tau." *Journal of the American Statistical Association*, 100, 484-492.

[74] McFadden, J. A. (1962). "On the lengths of intervals in a stationary point process." *Journal of the Royal Statistical Society. Series B*, 24, 364–382.

[75] Miyata, Y. (2004). "Fully Exponential Laplace Approximation Using Asymptotic Modes." *JASA*, 99, 1037-1049.

[76] Neyman, J. (1955). "Statistics; servant of all sciences". *Science*, 122, 401-406.

[77] Neyman, J. (1969). "Bias in Surveys Due to Nonresponse." In *New Developments in Survey Sampling*. Edited by Johnson & Smith. Wiley.

[78] Nowell, C., Evans, M. A., and McDonald, L. (1988). "Length-Biased Sampling in Contingent Valuation Studies." *Land Economics*, 64, 367-371.

[79] Nowell, C. and Stanley, L. R. (1991). "Length-Biased Sampling in Mall Intercept Surveys." *Journal of Marketing Research*, 28, 475-479

[80] Oakes, D. and Dasu, T. (2003). "Inference for the Proportional Mean Residual Life Model" *Crossing boundaries: statistical essays in honor of Jack Hall*, IMS Lecture Notes Monogr. Ser., 43, Inst. Math. Statist., 105-115.

[81] Patil G. P. and Rao, C. R.(1978). "Weighted Distributions and Size-Biased Sampling with Applications to Wildlife Population adn Human Families." *Biometrics*, 34, 179-189

[82] Patil, G. P., Rao, C. R. and Zelen, M. (1988)."Weighted distributions." *Encyclopedia of Statistical Sciences* (eds. S. Kotz and N. L. Johnson), 9, 565-571, Wiley, New York.

[83] Qin, J. (1998). "Inferences for case-control and semiparametric two-sample density ratio models." *Biometrika*, 85, 619-630.

[84] Robbins, H. and Zhang, C. H. (1988). "Estimating a treatment effect under biased sampling." *Proc Natl. Acad. Sci. USA*, 85, 3670-3672.

[85] Sackett, D. L. (1979) "Bias in analytic research." *J. Chron. Dis.*, 32, 51-63.

[86] Schoenfeld, D. (1982). "Partial residuals for the proportional hazards regression model." *Biometrika*, 69, 239-241.

[87] Severini, T. A. and Wong, W. H. (1992). "Profile likelihood and conditionally parametric models." *The Annals of Statistics*, 20, 1768-1802.

[88] Shun, Z. and McCullagh, P. (1995). "Laplace Approximation in High Dimensional Integrals." *JRSS B*, 57, 749-760.

[89] Smith, T. M. F. (1993) "Populations and Selection: Limitations of Statistics." *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 156, 144-166

[90] Stern, Y., Tang, M., Albert, M. S., Brandt, J., Jacobs, D. M., Bell, K., Marder, K., Sano, M., Devanand, D., Albert, S. M. and Tsai, W. (1997). "Predicting time to nursing home care and death in individuals with Alzheimer disease." *Journal of the American Medical Association*, 277, 806–812.

[91] Terwilliger, J. D., Shannon, W. D., Lathrop, G. M., Nolan, J. P., Goldin, L. R. Chase, G. A. and Weeks, D. E. (1997). "True and False Positive Peaks in Genomewide Scans: Applications of Length-Biased Sampling to Linkage Mapping." *American Journal of Human Genetics*, 61, 430-438.

[92] 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.

[93] Tsodikov, A. (1998). "A proportional hazards model taking account of long-term survivors." *Biometrics*, 54, 1508-1516.

[94] Turnbull, B. W. (1976). "The empirical distribution function with arbitrarily grouped, censored, and truncated data." *Journal of the Royal Statistical Society. Series B*, 38, 290–295.

[95] Van der Laan, M. J. and Hubbard, A. E. (1998). "Locally efficient estimation of the survival distribution with right-censored data and covariates when collection of data is delayed." *Biometrika*, 85, 771-783.

[96] Vardi, Y. (1982). "Nonparametric estimation in the presence of length bias." *Annals of Statistics*, 10, 616-620.

[97] Vardi, Y. (1985). "Empirical distributions in selection bias models." *Annals of Statistics*, 13, 178-205.

[98] Vardi, Y. (1989). "Multiplicative Censoring, Renewal Processes, Deconvolution and Decreasing Density: Nonparametric Estimation." *Biometrika*, 76, 751-761.

[99] Vardi, Y. and Zhang, C.-H. (1992). "Large sample study of empirical distributions in a random-multiplicative censoring model." *Annals of Statistics*, 25, 1022-1039.

[100] Wang, M.-C. (1989). "A semi-parametric model for randomly truncated data." *Journal of the American Statistical Association*, 84, 742–748.

[101] Wang, M.-C. (1991). "Nonparametric Estimation from Cross-Sectional Survival Data." *JASA*, 86, 130-143.

[102] Wang, M.-C. (1996). "Hazards regression analysis for length-biased data." *Biometrika*, 83, 343-354.

[103] Wang, M.-C., Brookmeyer, R. and Jewell, N.P. (1993) "Statistical Models for Prevalent Cohort Data." *Biometrics*, 49, 130-143

[104] Wang, M.-C., Jewell, N. P. and Tsai, W.-Y. (1986). "Asymptotic properties of the product limit estimator under random truncation." *The Annals of Statistics*, 14, 1597-1605.

[105] Wicksell, S. D. (1925). "The Corpuscle Problem. A Mathematical Study of a Biometric Problem." *Biometrika*, 17, 84-99.

[106] Wicksell, S. D. (1926). "The Corpuscle Problem: Second Memoir: Case of Ellipsoidal Corpuscles." *Biometrika*, 18, 151-172.

[107] Wolfson, C., Wolfson, D. B., Asgharian, M., M'Lan, C. E., Østbye, T., Rockwood, K. and Hogan, D.B. (2001). "A Reevaluation of the Duration of Survival After the Onset of Dementia." *The New England Journal of Medicine*, 344, 1111-1116

[108] Wong, R. (1989). *Asymptotic Approximations of Integrals.* Academic Press.

[109] Zelen, M. (2004). "Forward and Backward Recurrence Times and Length Biased Sampling: Age Specific Models." *Lifetime Data Analysis*, 10, 325-334.

[110] Zelterman, D. (1993). "A semiparametric bootstrap technique for simulating extreme order statistics." *Journal of the American Statistical Association*, 88, 477-485.

[111] Zhou, H., Weaver M.A., Qin, J., Longnecker, M.P. and Wang, M.C. (2002) "A Semiparametric Empirical Likelihood Method for Data from an Outcome-Dependent Sampling Scheme with a Continuous Outcome." *Biometrics*, 58, 413-421.