Duration Data Analysis in Longitudinal Surveys

by

Christian Boudreau

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2003

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Considerable amounts of event history data are collected through longitudinal surveys. These surveys have many particularities or features that are the results of the dynamic nature of the population under study and of the fact that data collected through longitudinal surveys involve the use of complex survey designs, with clustering and stratification. These particularities include: attrition, seam-effect, censoring, left-truncation and complications in the variance estimation due to the use of complex survey designs. This thesis focuses on the last two points.

Statistical methods based on the stratified Cox proportional hazards model that account for intra-cluster dependence, when the sampling design is uninformative, are proposed. This is achieved using the theory of estimating equations in conjunction with empirical process theory. Issues concerning analytic inference from survey data and the use of weighted versus unweighted procedures are also discussed. The proposed methodology is applied to data from the U.S. Survey of Income and Program Participation (SIPP) and data from the Canadian Survey of Labour and Income Dynamics (SLID).

Finally, different statistical methods for handling left-truncated sojourns are explored and compared. These include the conditional partial likelihood and other methods, based on the Exponential or the Weibull distributions.

Acknowledgments

I am grateful to Dr. Jerald F. Lawless, my supervisor, for his guidance, insight, patience and encouragements. I am also thankful to Dr. Mary E. Thompson for being my surrogate supervisor during Dr. Lawless' sabbatical, and for her help and guidance on what evolved to be chapter 3 of this thesis. I would like to acknowledge Dr. Jiahua Chen for his help with the proofs of section 3.2.

I wish to thank my thesis committee: Dr. Mary E. Thompson, Dr. Richard J. Cook, Dr. John Goyder and Dr. Danyu Lin (The University of North Carolina) for their dedication in reviewing my thesis, and for their helpful comments and suggestions.

I would also like to thank Dr. Pat Newcombe for her assistance at the South-Western Ontario Research Data Centre (SWORDC), where the analyses of section 4.2 were done.

My deepest gratitude goes to my mom and dad for their support, understanding, endless patience, encouragements and help. In particular, to my dad for his numerous editorial comments, and to my mom for her unconditional help when it was most needed and for always being there for me.

My thanks also go to all my friends who helped create wonderful memories of this journey. Those that were there at the beginning of this journey: Julia, Thierry, Laura, Monica, Jennifer and Robin; those that were there in the middle: Francis and the Mexican contingent (Ligia, Tulio, Paco, Norberto, Sandra and Patricia); and those that were there closer to the end: Cody, Marc, Aurélie and Denise. Special thanks to Diego Hernàndez and Dr. Ken Seng Tan who were there throughout this journey.

This research was supported by scholarships from the *Natural Sciences and Engineering Research Council of Canada* (NSERC) and by *Le Fonds pour la Formation de Chercheurs et l'Aide à la Recherche du Québec* (FCAR). I am indebted to both for allowing me to focus my energies on completing my doctorate.

To all thank you for helping me accomplish this dream.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

There is currently great interest in using longitudinal, panel, or cohort surveys to understand the different events that individuals experience over time. Examples of such events include marriage, divorce, fertility, spells of unemployment, spells on different social programs, etc. This information is referred to as the life history process of the individual. Event history analysis deals with the modelling and the analysis of such processes. Survival analysis is both a special case and a cornerstone of event history analysis. The distinction is that the former only deals with the modeling of data arising from observing individuals from some starting point until the occurrence of a specific event or end point, and not with the entire history processes of individuals.

Longitudinal surveys have many particularities or features not found in other types of surveys of finite population or super-population, as well as in classical analyses of independent observations. These features are mainly the result of the dynamic nature of the population under study and that data collected through longitudinal surveys generally involve the use of complex survey designs, with clustering and stratification. These features, which are discussed further in chapter 2, include: attrition, seam effect, censoring, left-truncation and complications in the variance estimation due to the use of complex survey designs. This thesis focuses on the last two points. Statistical methods that account for intra-cluster dependence, when the sampling design is uninformative, are proposed in chapter 3 and exemplified in chapter 4. The first part of chapter 5 contains a discussion on the conditional partial likelihood, a generalization of the Cox partial likelihood that allows the inclusion of left-truncated sojourns, in the context of longitudinal surveys. In addition, different statistical models, based on the Exponential or the Weibull distributions that allow for the inclusion of left-truncated observations, are proposed and compared in chapter 5. Finally, chapter 6 concludes this thesis by summarizing the main results and identifying areas for future research.

The rest of chapter 1 is organized as follows: section 1.1 presents three examples of longitudinal surveys that will be found throughout this thesis; section 1.2 contains the basis of event history analysis; section 1.3 introduces the martingale framework; and section 1.4 discusses the basis of left-truncation.

1.1 Examples

This section introduces three longitudinal surveys that will be used as examples in this thesis. The Survey of Labour and Income Dynamics (SLID) is the subject of section 1.1.1, the Survey of Income and Program Participation (SIPP) of section 1.1.2 and the Panel Study of Income Dynamics (PSID) of section 1.1.3. They will be used to exemplify the particularities of longitudinal surveys that will be discussed in chapter 2. In addition, the methods proposed in chapter 3 will be applied to SIPP and SLID data in chapter 4. These surveys illustrate the complexity of large scale longitudinal studies and the many goals they have to fulfill under numerous constraints.

1.1.1 Survey of Labour and Income Dynamics

Building on the experience acquired with the Labour Market Activity Survey (LMAS) in the 1980's, Statistics Canada launched the Survey of Labour and Income Dynamics (SLID) in 1992–1993. The goal was to fill the need for longitudinal information on individuals and families. Thus, SLID was designed to capture changes in the economic well-being of individuals and families over time. It focuses on income dynamics (e.g., transitions into and out of poverty), labour market dynamics (e.g, spells of unemployment and transitions from school to work) and family dynamics (mainly their influences on the economic well-being).

As shown in figure 1.1, the 1st panel started on January 1, 1993 and new panels are introduced every 3 years (e.g., the 2nd panel started on January 1, 1996, the 3rd panel on January 1, 1999 and so on). Since participants are followed over a 6 year period (e.g., up to December 31, 1998 for the 1st panel), successive panels have a 3 year overlap. The choice of following participants for 6 years was influenced by costs, design considerations as well as limiting respondent burden. The target population for SLID is all persons living in Canada, excluding the people residing in Yukon, Northwest Territories or Nunavut, persons living in institutions (same definition as the one used for the Census) or on Indian reserves and full-time member of the Canadian Armed Forces living in barracks. At the beginning of each panel, an initial sample is drawn from the Labour Force Survey (LFS). That LFS sample is based on a stratified multi-stage sampling design. More precisely, the stratification is done according to the following scheme. First, each province is divided into LFS economic regions. Second, these regions are further divided into one or more "urban" areas and one "rural" area, where the rural area is defined as the area of the stratum not covered by urban areas. Finally, each urban area is further subdivided into strata, which have similar socioeconomic characteristics. The primary sampling units (PSU's) are clusters formed by identifying groups of dwellings (e.g., city blocks) within each of these strata. A random sample of households is then selected within each PSU. In the case of the 1st panel, this resulted in a longitudinal sample of about 15,000 households or about 30,000 participants aged 16 and over in January 1993. Since all individuals living in a sampled household are followed by SLID, it is important to account for the resulting correlation in any statistical procedures. One would also expect some degree of correlation between households belonging to the same PSU and/or the same stratum.

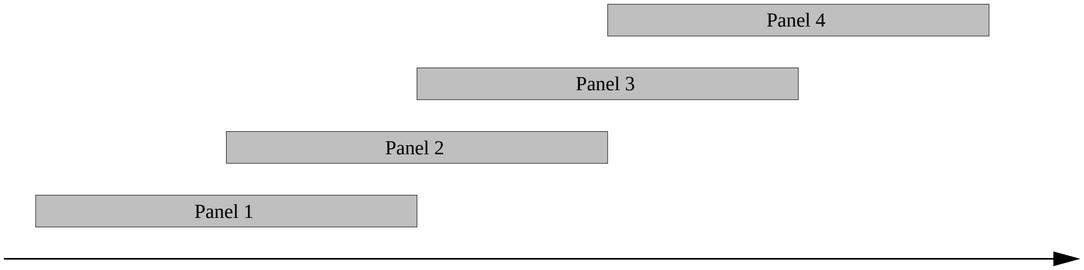

Figure 1.1: SLID panels design.

Like most Statistics Canada surveys, participation in SLID is not mandatory under the Statistics Act, but on a voluntary basis. Many efforts are taken to keep the response rate as high as possible and limit attrition; see section 2.3 for further discussion. The measures taken to improve the response rate and data quality include computer assisted interviews (CAI), and reducing respondent burden, by limiting the number of interviews to a total of 12 (13 for the 1st panel) over the 6 year period. CAI helps in insuring consistency between interviews and between labour and income information. It also allows feeding back past reported information. This latter point is important because it helps in reducing seam-effect problems, which can arise with long recall periods; see section 2.1. There are other advantages to CAI, such as on-line editing of information and dependent interviews (i.e., questionnaires that vary according to the answers given by respondents). The interview process is divided into two parts: labour and income. The labour part is conducted every January and collects information on events that arose over the previous calendar year. This includes topics such as employment and unemployment, family relationships, education, etc. The income part is conducted every May and, like the labour interview, collects information regarding the previous calendar year. It is done in May to take advantage that most respondents filed their income tax report in April and should be more familiar with their financial records. Note that this interview is not required if the respondent chooses to give Statistics Canada permission to access his/her income tax report through Revenue Canada. For the 1st panel an extra interview was done in January 1993 to collect background information. From panel 2 onward that interview is done together with the first labour interview. Participants under 15 years of age are not interviewed, but some of their information is collected from other members of the household. The use of proxy is common in SLID and is not reserved to children. This explains some of the missing information in SLID data.

The interview format described in the previous paragraph does not account for the dynamic or ever-changing nature of SLID. In addition to people being lost to follow up, households may split up, new people may join existing households, some participants may move in and out of the target population described previously, etc. To account for these factors, the interview format must be flexible. In SLID this was done according to the following rules. If a household splits up (e.g., through divorce or older children moving out) all branches are followed. New people joining existing households (e.g., through marriage or newborns) are included in the survey and interviewed. However, they receive a longitudinal sampling weight of zero and are only followed as long as they remain in that household. They are referred to as "cohabitants" in SLID. These cohabitants or non-sampled individuals are usually referred to as top-ups in longitudinal surveys; see section 2.4. However, the word top-up sample in SLID refers to a different group of participants; see page 8. People leaving the target population, as described previously, are followed as long as possible. The amount of information collected on these people will depend on the reason they are no longer part of the target population. For example, people moving to Yukon, Northwest Territories, Nunavut or the continental United States will be interviewed as they were before. However, people moving to other countries or to an institution will only be traced so that the interviewing process can resume if they return to the target population. These people are referred to as "movers" in SLID.

Information on more than 1,000 variables is collected on each participant over the 6 year period they are followed. These are grouped into four broad categories: personal characteristics, education, income, wealth and labour. These categories cover 14 themes (see figure 1.2), which are:

1. Demographics (e.g., age, sex and marital status);

2. Ethnocultural characteristics (e.g., mother tongue, immigrant status, ethnic background, parents' education and visible minority status);

3. Activity limitations (e.g., screening questions on conditions limiting the amount or kind of activities an individual can do at home and at work);

4. Information of person's children (e.g., how many children he/she has, and the age the individual was when his/her first child was born);

5. Geography (e.g., individual's place of residence, if he/she moved in the last year and if he/she is still part of the same household as last year);

6. Family and household characteristics (e.g., number of persons living in the household, type of dwelling and income sources of the economic family[1]);

7. Labour market activity patterns (e.g., number of jobs held, total hours worked at all jobs, total hours paid, absences from work and wages);

8. Work experience (e.g., number of years of work experience at full-time or equivalent);

9. Jobless periods (e.g., dates and durations of jobless spells, and whether or not the person looked for work during the jobless spell);

10. Job characteristics (e.g., dates and durations of job spells, work schedule, hourly wage, total earning and overtime);

11. Income sources (e.g., wages and salaries, investment income, total income and low income status);

[1] Economic family is a broad definition of family; it includes all persons sharing a dwelling and related by blood, marriage, common-law or adoption.

12. Employment insurance, worker's compensation and social assistance (e.g., which months these government transfers were received and what amounts were received);

13. Educational activity (e.g., student status, type of school attended and whether the respondent obtained a degree that year);

14. Educational attainment (e.g., years of schooling, indicator of high-school completion and various information on postsecondary education).

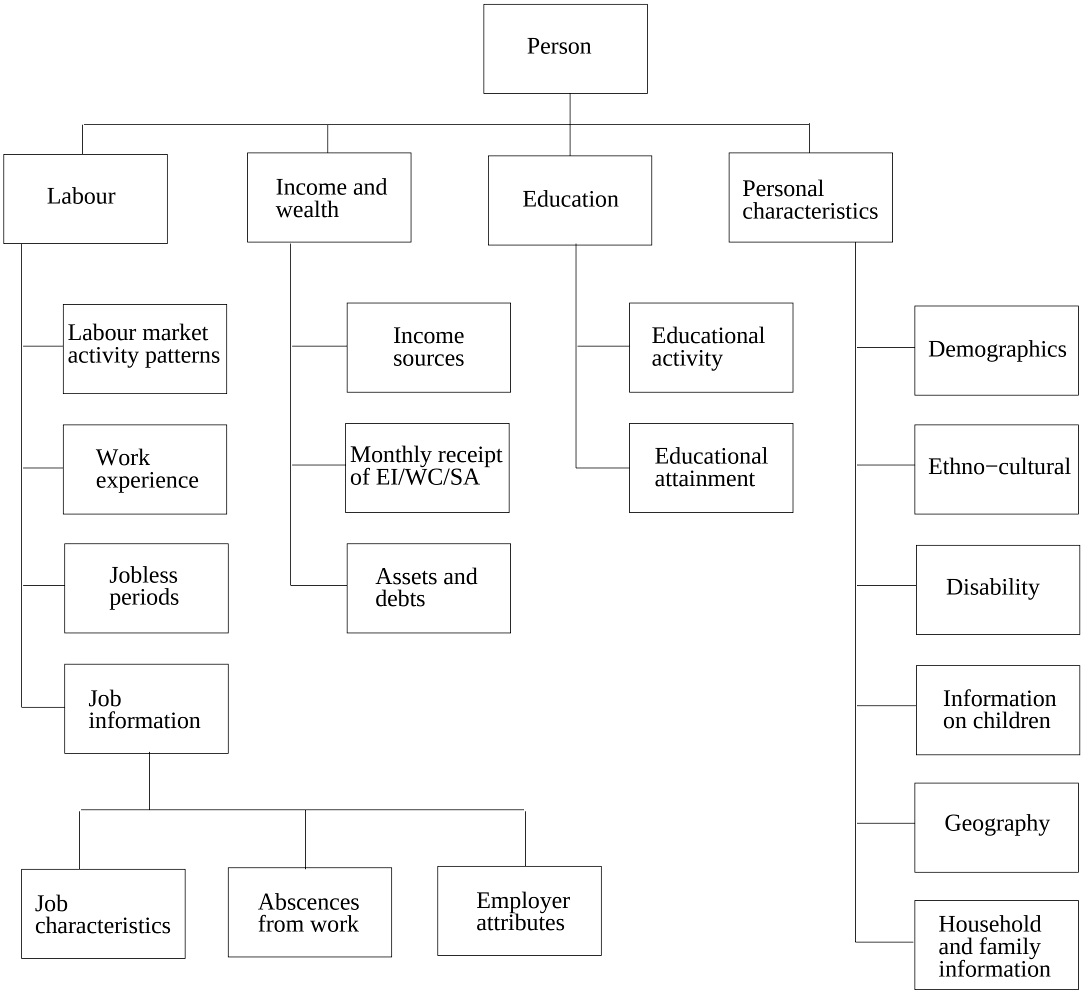

Figure 1.2: SLID categories and themes.

Theme 9, jobless periods, will be used in section 4.2 to illustrate the methods proposed in chapter 3. This paragraphs gives further information on the data collected on jobless spells. In SLID, a jobless spell refers to a period in which a person had no attachment to any employer. Therefore, seasonal layoff or other type of layoff where the person is expected to return to work with the same employer are treated as absence from work. There is thus a distinction between being jobless and unemployed. Dates and durations (in weeks) of jobless spells are derived from the start and end dates of jobs held during the year. Information for the other themes (e.g., sex, age, number of children being raised, years of schooling and employment insurance payments) is also available for each jobless spell. Whether or not the person looked for work and, if not, why is valuable information that can be used to analyse discouragement, for example.

SLID is first and foremost a longitudinal survey. However, it also has a cross-sectional component. To maintain cross-sectional representativity, the longitudinal sample is augmented each year by a randomly selected sample of about 9,000 households. These yearly samples are referred to as top-ups in SLID, and a new one is drawn each year replacing the previous one. In accordance, cross-sectional weights, which differ from longitudinal weights, are computed each year. These weights are appropriate to calculate estimates based on data from a single year. However, the longitudinal weights reflect the population at the time the panel participants were selected and are the most appropriate for longitudinal analyses. Only longitudinal respondents are assigned a non-zero longitudinal weight. The longitudinal weights are recalculated each year to account for non-response and attrition; see section 2.6.

Documentation on SLID, in both paper and electronic formats, is available from Statistics Canada. The Survey of Labour and Income Dynamics Microdata User's Guide, which also contains the data dictionary, is a comprehensive source of information. There is also an extensive working paper series covering various research themes, such as employment and unemployment dynamics, life events, family changes, educational achievement, life cycle labour market transitions, job quality and dynamics of low income. These working papers and the SLID User's Guide (Statistics Canada (1997)) are available at the following web sites:

- `www.statcan.ca/english/freepub/75M0001GIE/free.htm` ;
- `www.statcan.ca/english/studies/` ;
- `www.statcan.ca/cgi-bin/downpub/listpub.cgi?catno=75F0002MIE` ;
- `www.statcan.ca/cgi-bin/downpub/listpub.cgi?catno=11F0019MIE` .

SLID data can be obtained through public-use microdata files or, as it was done in this thesis, through a Statistics Canada Research Data Centre. In the first case, confidentiality of individual records is mainly ensured by deleting, grouping, rounding or adding random perturbations to some variables. In the second case, individual records are available, but only to researchers sworn under the Canadian Statistics Act.

1.1.2 Survey of Income and Program Participation

A survey that shares many similarities with SLID is the Survey of Income and Program Participation (SIPP) conducted by the U.S. Census Bureau. SIPP is a continuing panel survey of the U.S. population, with a new panel being introduced each year. Like SLID, SIPP builds on the knowledge acquired through a previous survey, the March Income Supplement of the Current Population Survey (CPS). However, CPS is a cross-sectional survey and is primarily designed to obtain information on employment and unemployment. The longitudinal features of SIPP allows for much broader analyses than the CPS, such as participations to various social programs, changes in poverty status and their associated events. It provides rich and valuable data on sources and amounts of income, labor force information, program participation and eligibility. The effects of federal and state programs on the economic well-being of families and individuals can also be evaluated through SIPP. This allows for a more efficient design and better targeting of various transfer programs. One aim of this survey is to estimate the duration of spells of participation in different social programs and of spells without health insurance. The transfer programs studied are food stamps, Aid to Families with Dependent Children (AFDC), general assistance, Social Security, Federal Supplemental Security Income (SSI), Veterans compensation or pensions, State unemployment compensation and Women, Infants & Children Nutrition program (WIC).

A national sample of U.S. households (see Citro et al. (1986) for the definition of "household" adopted in SIPP) is introduced every year and followed over a 32 month period (some panels, like the 1996 panel, have a followup period of 4 years). Hence, two successive panels overlap for 20 months. Interviews are conducted at 4 months intervals; resulting in a total of 8 interviews for the usual 32 month followup period. The short recall period is designed to reduce respondents' difficulties in recalling specific information and, hopefully, increase accuracy.

The first SIPP panel was introduced in 1984 and started with a sample of 19,878 households. The newer 1996 panel consists of 36,700 sampled households. This thesis focuses on the 1987 panel, which consits of a sample of approximately 12,000 households or 24,428 individuals with strictly positive longitudinal weights. The target population for SIPP is all non-institutionalized U.S. civilians. There is a greater focus on participants aged 15 years and over, but information on younger participants is collected through other members of the household or family. At the beginning of each SIPP panel, a national sample of households is drawn using a stratified-multistage sampling design similar to the one used for the CPS. Each household initially selected is called a sample unit and is assigned a Sample Unit ID (SUID). A household is composed of a primary family (which is the equivalent of SLID economic family; see footnote on page 6) and, possibly, of sub-families (e.g., daughter and husband living with parents). All members that were part of a selected household at the beginning of a panel are assigned a strictly positive longitudinal weight and followed throughout the observation window. This last point is common to both SIPP and SLID. Therefore, individuals belonging to the same SUID are expected to be correlated.

To facilitate field procedures and to spread out the work load for interviewers, each panel is randomly divided into four groups of approximately equal size, called "rotation groups". One rotation group is interviewed each month successively, resulting in coverage of the entire sample in 4 months. Each round of interviews is referred to as a wave. The data collected in each wave include labor force status, program participations and income during the 4 calendar months preceding the interview, referred to as the "reference period" in SIPP. As mentioned previously, a short reference period was chosen to improve data accuracy and reduce recall bias, such as the seam-effect. Unfortunately, seam-effect errors are widespread in SIPP. Hence, the number of reported transitions is greater between months for which the data were collected in different waves than between months for which the data were collected in the same wave. See Coder et al. (1987), Kalton & Miller (1991) and section 2.1 for further discussion on the seam-effect in SIPP and in other longitudinal surveys.

Interviews are conducted by personal visits and by decentralized telephones. This allows interviews of persons who will not or can not participate otherwise and to obtain information which would, most likely, be missing if interviews were done in other ways. The 1996 SIPP panel saw the introduction of computer assisted interviews (CAI). All participants aged 15 years and over are interviewed by self-response, if possible. Proxy response is permitted when participants are not available and to collect information on children who live in the same household. Non-sampled individuals joining existing households (e.g., through marriage) are followed as long as they reside with an original sampled person. These individuals are referred to as top-ups; see section 2.4 for further information. As it was done in SLID, these top-up individuals are assigned longitudinal weights of zero, and thus excluded from weighted longitudinal SIPP data analyses. A great deal of effort is taken to follow movers. These include contacting neighbors, employers and postal supervisors. For example, the Census Bureau mails advance letters to respondents before each interview; if the respondent no longer lives at the address, the post office is requested to provide a forwarding address. Original sampled persons institutionalized after the first interview, are tracked but not interviewed. The interview process resumes if they reenter the target population. Persons moving abroad or into Armed Forces barracks are not followed. These procedures for following movers were introduced in wave 6 of the 1984 panel and wave 2 of the 1985 panel. Before, movers were not followed unless they moved in with other original sampled persons still in the survey. For more information about tracking persons over time see Jean & McArthur (1987).

SIPP contains information on about 230 distinct variables. Sex, age and race questions are asked at the first interview. A second group of variables are collected through questions that vary from interview to interview. The third group corresponds to questions that are asked at every interview, such as highest grade of school achieved and in which U.S. state they reside. These are called core questions and are classified into five sections:

1. Labor force and recipiency: Over 30 questions are asked to each participant. The numbers of weeks worked and weeks without a job during the reference period are determined. Further questions are asked to participants that have experienced one or more jobless spells during the reference period to determine if they were part of the labor force[2] or not. Participation in various transfer programs (see page 9 for the list of programs) during the reference period is also established. Finally, miscellaneous sources of income are listed, and information about Medicare, Medicaid and private health care coverage.

2. Earnings and employment: For each job identified in section 1, the type of business or industry, the kind of work done, total wage or pay and main activities or duties of the participant are established through a list of questions. Self-employed individuals go through a similar process.

3. Amounts of income received: This section gathers the information about amounts received from programs and other income sources identified in section 1.

[2] Individuals without a job, not on layoff and not looking for work are not considered to be part of the labor force in SIPP.

4. Program questions. The questions of this section are asked to determine if the household participates in energy assistance and in subsidized school lunch or breakfast programs.

5. Missing wave: This section was introduced in wave 4 of the 1984 panel. It allows the interviewer to collect information on participants that could not be interviewed in the preceding wave, but who were previously interviewed.

The 1987 panel will be the focus of section 4.1. More precisely, the effects of sex, age, race, education, income and other variables on the duration of spells on the food stamps program (section 4.1.1) and of spells without health insurance (section 4.1.2) will be explored. This will also serve as an example for the methods proposed in chapter 3. Participation in the food stamps program is established by asking each individual if he/she was covered by the program for each of the 32 months. Lack of health insurance coverage is established by asking each individual three coverage questions for each of the 32 months. These questions determined whether an individual was covered through health insurance in his/her own name, through insurance in someone else's name or through their employer. Lack of health insurance coverage is defined as a negative answer to all of the coverage questions. The sample considered contains 35,944 individuals divided into 46 strata corresponding to the 50 states plus the District of Columbia; some states are grouped together. Clusters are formed using SUID identifier, which generally corresponds to a household.

Like SLID, SIPP was primarily designed as a longitudinal survey. Computer files, containing all the data on core questions, are also produced for cross-sectional analyses. A vast amount of information is available on SIPP. One source is the Survey of Income and Program Participation Users' Guide (U.S. Census Bureau (2001)), published in both paper and electronic formats. The U.S. Census Bureau web site provides basic information on SIPP at `www.sipp.census.gov/sipp/`. Reports and working papers are also available from that web site, see `www.sipp.census.gov/sipp/pubsmain.htm`.

1.1.3 Panel Study of Income Dynamics

The Panel Study of Income Dynamics (PSID) is a joint project between the U.S. Census Bureau and the Institute for Social Research (ISR), University of Michigan. With a follow-up period of 20 years, it has a unique position among longitudinal surveys. In 1968, a sample of 4,802 households/families was randomly selected. Since one of the main goals of PSID was the study of poverty and its determinants, low-income households were over sampled. This portion of the sample became known as the Survey of Economic Opportunity Sample (SEO). It is composed of 1,872 families, which had a total income equal or less than twice the poverty level for the corresponding family size in 1966. The follow-up for these families was done by the U.S. Census Bureau. The other portion of the sample was selected from the master sampling frame of the Survey Research Center, one of the four research centers of the ISR. These 2,930 families became known as the Survey Research Center Sample (SRC) and follow-up was handled by the ISR.

Interviews were conducted yearly for a total of 21 waves by the end of the survey in 1988. Similar features as the ones observed in SLID and SIPP are also present in PSID. These include loss to follow-up or attrition and addition of new people joining existing families or top-ups. Only participants that were part of the original 1968 sample and newborns (if both parents were part of the original sample) were assigned non-zero weights. Further information can be found in the User Guide to the Panel Study of Income Dynamic (Institute for Social Research (1984)) and in Lillard (1989).

The last three sections of chapter 1 discuss statistical and mathematical background material that will be used throughout this thesis.

1.2 Basis of event history analysis

Event history analysis is based on counting processes, and can be viewed from two different frameworks: the event occurrence and multi-state frameworks. Both are mathematically equivalent, but depending on the problem at hand one may be preferred to the other for descriptive or explanatory purposes. If the interest is in modeling the number of times an individual experiences a certain event (e.g., number of epileptic attacks suffered by a patient), then the event occurrence framework is more practical. On the other hand, if the interest is in modeling transitions between states or duration of sojourns (e.g., duration of jobless spells), then the multi-state framework is preferred.

To formalize the multi-state framework, let

$$Y_i(t) = \text{state occupied by individual } i \text{ at time } t \ , \tag{1.1}$$

where $i = 1, \ldots, n$ and $Y_i(t)$ takes values in $\{1, 2, \ldots, K\}$. That is, each individual must occupy a unique state among the finite set $\{1, 2, \ldots, K\}$ at any given time. Therefore, $\{Y_i(t); t \geq 0\}$ keeps track of the different transitions that occur to an individual over time. The information recorded includes the transitions between different states, as well as the time at which these transitions occurred. The event occurrence framework also keeps track of the same information, recording it in a slightly different manner.

Accordingly, in the event occurrence framework, let

$$N_{ij}(t) = \text{number of occurrences of type } j \text{ event for individual } i \text{ up to time } t \ , \tag{1.2}$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, J$; there are J possible types of events. Again, the two frameworks are equivalent since state transitions can be viewed as events and vice-versa; that is, a type j event corresponds to a $k \rightarrow \ell$ transition, say.

Stochastic models for event history analysis are based on counting processes and are generally defined in terms of intensity functions as follows. First, let $\mathrm{H}_i(t^-)$ denote the history of all events and covariates relevant to individual i up to, but not including, time t (e.g., $\mathrm{H}_i(t^-) = \{Y_i(u); 0 \leq u < t\}$). For multi-state models, these functions are called transition intensity functions and are defined as

$$\lambda_{ik\ell}(t \mid \mathrm{H}_i(t^-)) = \lim_{\Delta t \rightarrow 0} \frac{\Pr\{Y_i(t + \Delta t) = \ell \mid \mathrm{H}_i(t^-), Y_i(t^-) = k\}}{\Delta t} \ , \tag{1.3}$$

where $k \neq \ell$ and both k and ℓ range over $\{1, 2, \ldots, K\}$. In short, (1.3) means that, given its prior history (i.e., $\mathrm{H}_i(t^-)$ which includes the fact that he/she is in state k at time t^-), the conditional probability of individual i making a transition to the new state ℓ by time $t+\Delta t$ is approximately $\lambda_{ik\ell}(t \mid \mathrm{H}_i(t^-))\,\Delta t$ for small Δt. Also, the probability of more than one transition in the short interval $[t, t+\Delta t)$ is of order $\mathrm{o}(\Delta t)$.

Likewise, in the event occurrence framework, event intensity functions are defined as

$$\lambda_{ij}(t \mid \mathrm{H}_i(t^-)) = \lim_{\Delta t \to 0} \frac{\Pr\{N_{ij}[t, t+\Delta t) = 1 \mid \mathrm{H}_i(t^-)\}}{\Delta t}\,, \tag{1.4}$$

where $N_{ij}[s,t) = N_{ij}(t^-) - N_{ij}(s^-)$ is the number of type j events in the interval $[s,t)$ and $j = 1, \ldots, J$. In other words, (1.4) implies that, given its prior history, the conditional probability of individual i experiencing a type j event in $[t, t+\Delta t)$ is approximately $\lambda_{ij}(t \mid \mathrm{H}_i(t^-))\,\Delta t$ for small Δt.

Intensity functions play a fundamental role since other quantities in event history analysis can be expressed in terms of (1.3), in the multi-state framework, or of (1.4), in the event occurrence framework. In particular, the probability of no exit from state k by time $t+s$, given the history up to time t^-, is given by

$$\Pr\{\text{No exit from state } k \text{ by } t+s \mid \mathrm{H}_i(t^-)\} = \exp\left\{-\int_t^{t+s} \sum_{\ell \neq k} \lambda_{ik\ell}(u \mid \mathrm{H}_i(u^-))\,du\right\}\,. \tag{1.5}$$

Similarly, for the event occurrence framework,

$$\Pr\{\text{No event over } [t, t+s) \mid \mathrm{H}_i(t^-)\} = \exp\left\{-\int_t^{t+s} \sum_{j=1}^{J} \lambda_{ij}(u \mid \mathrm{H}_i(u^-))\,du\right\}\,. \tag{1.6}$$

Combining the previous equations, it is possible to express the probability density of an individual's history process up to time t (i.e., $\{Y_i(u); 0 \leq u \leq t\}$). For example, in the event occurrence framework, if individual i experienced m events $j_1, j_2, \ldots, j_m$ at times

$t_1 < t_2 < \ldots < t_m$, then the corresponding density function is given by

$$\begin{aligned}
&\Pr\{m \text{ events in } [0,t] \text{ of types } j_1,\ldots,j_m \text{ at times } t_1 < \ldots < t_m \mid \mathrm{H}_i(0)\} \\
&= \prod_{r=1}^{m} \lambda_{ij_r}(t_r \mid \mathrm{H}_i(t_r^-)) \, \exp\left\{ -\int_0^t \sum_{j=1}^{J} \lambda_{ij}(u \mid \mathrm{H}_i(u^-))\, du \right\} \\
&= \prod_{r=1}^{m}\prod_{j=1}^{J} \lambda_{ij}(t_r \mid \mathrm{H}_i(t_r^-))^{\delta_{ij}(t_r)} \, \exp\left\{ -\int_0^t \sum_{j=1}^{J} \lambda_{ij}(u \mid \mathrm{H}_i(u^-))\, du \right\} , \qquad (1.7)
\end{aligned}$$

where

$$\delta_{ij}(t) = \begin{cases} 1 & \text{if individual } i \text{ experiences a type } j \text{ event at time } t \\ 0 & \text{otherwise .} \end{cases}$$

Likewise, in the multi-state framework,

$$\begin{aligned}
&\Pr\{m \text{ transitions in } [0,t] \text{ at times } t_1 < \ldots < t_m \mid \mathrm{H}_i(0)\} = \\
&\prod_{r=1}^{m}\prod_{k=1}^{K}\prod_{\ell\neq k} \lambda_{ik\ell}(t_r \mid \mathrm{H}_i(t_r^-))^{\delta_{ik\ell}(t_r)} \, \exp\left\{ -\int_0^t \sum_{\ell'\neq k} Y_{ik}(u)\, \lambda_{ik\ell'}(u \mid \mathrm{H}_i(u^-))\, du \right\} , \quad (1.8)
\end{aligned}$$

where

$$\delta_{ik\ell}(t) = \begin{cases} 1 & \text{if individual } i \text{ experiences a } k \rightarrow \ell \text{ transition at time } t \\ 0 & \text{otherwise} \end{cases}$$

and

$$Y_{ik}(t) = I(Y_i(t) = k) = \begin{cases} 1 & \text{if individual } i \text{ is in state } k \text{ at time } t \\ 0 & \text{otherwise .} \end{cases}$$

1.2.1 Event history models with covariates

In many longitudinal studies the effects of various covariates on transition or event intensities or probabilities are also of interest. For example, in studying jobless spells, the influences of the following covariates may be of interest: sex, age and level of education.

Covariates can be classified into two categories: external and internal covariates. In the first case, $x_i(t)$ take values that are, conditional on $\mathrm{H}_i(t^-)$, independent of the event history process $\{Y_i(t); t \geq 0\}$. For example, $x_i(t)$ is the amount of pollen in the prediction of the frequency of asthma attacks in children. In other words, the covariates do not directly carry information on the event history process of the corresponding individual. Another example of external covariates is fixed covariates, whose values are measured at the start of the study and fixed for the duration of the study (e.g., sex and race). Formally, $x_i(t)$ is a fixed covariates if $x_i(t) = x_i \; \forall \; t \geq 0$. On the other hand, internal covariates carry information about the history process of the corresponding individual. An example is a covariate that record the white blood cell count of a patient suffering from cancer. In this case $x_i(t)$ is directly related to the survival of patient i. Analysis with internal covariates involves using a joint model for the covariates and the event processes, and statistical inference is carried out using partial likelihood. Internal covariates are not considered further in this thesis. See Kalbfleisch & Prentice (2002), section 6.3, or Lawless (2003), pages 35–36, for additional discussion on external and internal covariates.

The theory of counting processes, on which event history analysis is based, is rich and flexible, and covariates are easily handled and incorporated in the previous models. Let $x_i(t)$ be the vector of all fixed and time varying covariates associated with individual i at time t. The definition of $\mathrm{H}_i(t^-)$, given on page 15, already includes the history of the covariates up to time t^-. By conditioning on $\mathrm{H}_i(t^-) = \{Y_i(u), x_i(u); 0 \leq u < t\}$ and $x_i(t)$ instead of only $\mathrm{H}_i(t^-)$, (1.3) to (1.8) can handle covariates without further modification.

1.2.2 Survival analysis

As mentioned previously, survival analysis can be viewed as a special case of event history analysis. A survival model can be considered as a transitional model with two states: 1 — alive and 2 — dead. The only transition allowed is from state 1 to state 2, which is an absorbing state, meaning that when an individual reaches state 2 its life history process has ended. Similarly, for the event occurrence framework, survival models can be regarded as models where only one type of event is possible, death; that is, $J = 1$. This is best illustrated by figure 1.3.

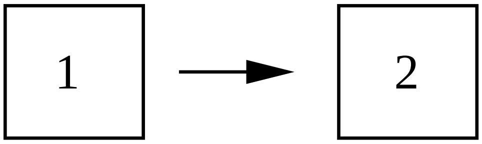

Figure 1.3: Survival model.

For survival models, the only relevant prior information contained in $\mathrm{H}_i(t^-)$ is if the individual is alive or dead at time t^- and the covariates history. Therefore, (1.3) simplifies and is called the hazard function,

$$\lambda_i(t \mid X_i(t)) = \lim_{\Delta t \to 0} \frac{\Pr\{t \leq T_i < t + \Delta t \mid T_i \geq t, X_i(t)\}}{\Delta t} \; , \tag{1.9}$$

where T_i is the time spent in state 1 and $X_i(t) = \{x_i(u); 0 \leq u \leq t\}$. If only fixed covariates are present (i.e., $x_i(t) = x_i \; \forall \; t \geq 0$), then (1.5) also simplifies and is called the survivor function,

$$S_i(t \mid x_i) = \Pr\{T_i \geq t \mid x_i\} = \exp\left\{ - \int_0^t \lambda_i(u \mid x_i) \, du \right\} \; . \tag{1.10}$$

Survival models can also be viewed as a building block of event history as they are at the basis of semi-Markov or duration models that will be discussed in the next section.

1.2.3 Semi-Markov or sojourn duration models

Intensity probabilities (1.3) and (1.4) can depend on such features as the states previously occupied, the number of previous events, the time since the current state was entered, covariates, etc. Such models can rapidly become complicated and require large datasets to carry out statistical inference. Therefore, simpler models such as Markov, semi-Markov or sojourn duration models are generally preferred. Markov models make the assumption that (1.3) depends on $\mathrm{H}_i(t^-)$ only through the state currently occupied, $Y_i(t^-)$, and the covariates $x_i(t)$.

Semi-Markov or sojourn duration models make the assumption that (1.3) depends on $\mathrm{H}_i(t^-)$ only through the state currently occupied, the elapsed time since that transition occurred and the covariates. For such models, it is useful to introduce a second time origin. Hence, let t be the time elapsed since the last transition and this "clock" is reset to $t = 0$ at each transition. As before, the time elapsed since the origin of the individual life history process (e.g., birth) is recorded and will be denoted by s, with time origin at $s = 0$. However, that "clock" is never reset to $s = 0$. This is portrayed in figure 1.4. Therefore, (1.3) simplifies and can be rewritten as

$$\lambda_{ik\ell}(t \mid x_i(s)) = \lim_{\Delta t \to 0} \frac{\Pr\{Y_i(s + \Delta t) = \ell, t \leq T_{ik} < t + \Delta t \mid Y_i(s^-) = k, T_{ik} \geq t, x_i(s)\}}{\Delta t}\,, \quad (1.11)$$

for $k \neq \ell$ and where T_{ik} is the r.v. corresponding to the time elapsed since the current state k was last entered. Likewise, (1.5) can be rewritten as

$$\Pr\{\text{No exit from state } k \text{ after a spell of less than } t \text{ time units} \mid Y_i(s^-) = k, T_{ik} \geq t, X_i(\infty)\} = \exp\left\{-\int_0^t \sum_{\ell \neq k} \lambda_{ik\ell}(u \mid x_i(s - t + u))\, du\right\}\,, \quad (1.12)$$

where $X_i(\infty) = \{x_i(u); u \geq 0\}$. Similar simplifications arise for (1.4) and (1.6) in the event occurrence framework.

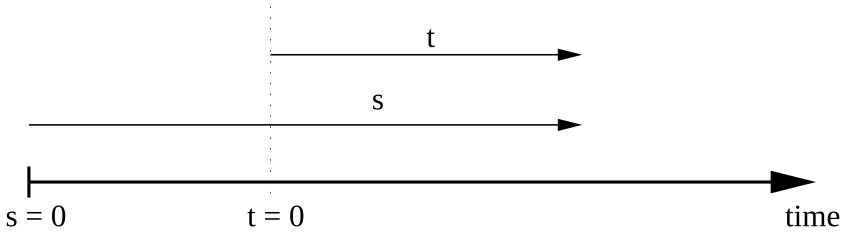

Figure 1.4: Time origins in semi-Markov models.

Sojourn duration or semi-Markov models are useful in many settings. The duration of a sojourn can be viewed as the time between two events, E_1 and E_2, say. Two or more duration models can be combined to yield more complex models. Classical examples include: three state progressive, illness-death and two state cyclical models. The first of these models is portrayed in figure 1.5. It has three states and the only possible transitions are $1 \rightarrow 2$ and $2 \rightarrow 3$. Duration of first marital unions is often studied using a three state progressive model. In that case, the three states are: 1 — never been married, 2 — in his/her first marriage and 3 — dissolution of that first marriage. That is, E_1 is the entry in first marital union and E_2 is the dissolution (by death or divorce) of that union. Even though both $1 \rightarrow 2$ and $2 \rightarrow 3$ transitions can be of interest, the focus is generally on the latter one, as well as the effects of possible covariates. If the duration of the first marital union does not depend on the age at marriage (i.e., time of entry to state 2), then (1.3) simplifies and the transition intensity function from state 2 to state 3 is given by (1.11), where s_i and t_i are, respectively, the age of individual i and how long his/her marital union lasted. In some settings, however, it might be required to allow dependence on the age at marriage to get a satisfactory model. The illness-death model is portrayed in figure 1.6 and was introduced to model disease progression. Like the three state progressive model, it has three states; often referred to as: 1 — alive and disease free, 2 — alive with disease and — 3 dead. In addition to $1 \rightarrow 2$ and $2 \rightarrow 3$ transitions, it allows for a direct $1 \rightarrow 3$ transition. The last example of sojourn duration models is the two state cyclical model portrayed in figure 1.7. It is used to model the time spent in one or both states in a model consisting of two alternating states and where the possible transitions are $1 \rightarrow 2$ and $2 \rightarrow 1$. Jobless spells (see sections 1.1.1 and 4.2.1), spells without health insurance (see sections 1.1.2 and 4.1.2), or spells between periods of depression in a patient suffering from bi-polar disorder or manic-depression are good examples of the use of this type of model.

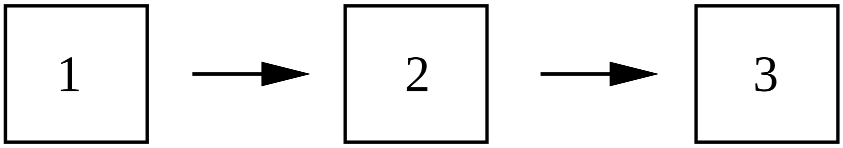

Figure 1.5: Three state progressive model.

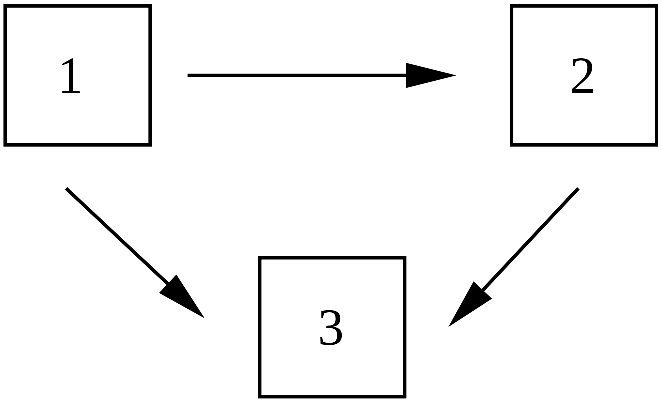

Figure 1.6: Illness-death model.

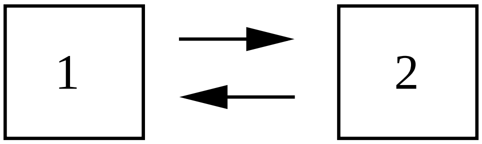

Figure 1.7: Two state cyclical model.

Additional event history models include the competing risk (or multiple modes of failure) process. This situation arises when an individual may die from different, but mutually exclusive, causes. Similarly, a piece of equipment may fail because of K different reasons. This process is portrayed by a model with $K+1$ states, one of them being alive and the K others being mutually exclusive modes of failure.

1.2.4 Observation window, censoring and truncation

Event history data usually arises from studying individuals, subjects or units over some time period or observation window. In many cases, the observation window is much shorter than the complete history process under study. This is particularly true for longitudinal surveys, where individuals are followed over a relatively small fraction of their life (e.g., 6 years for SLID and 28 months for SIPP). This is mostly due to cost, time and other constraints that investigators have. Loss to follow-up and a subject's refusal to participate further are other causes. Therefore, the variables $Y_i(t)$ or $N_{ij}(t)$ and covariates $x_i(t)$ are recorded only over the time interval $[\tau_{i0}, \tau_{i1}] \subseteq [0, \infty)$, where τ_{i0} and τ_{i1} are stopping times; see section 1.3.2. Note that the observation window can vary from subject to subject.

Possible scenarios of complete and incomplete event history data concerning the time interval between two events E_1 and E_2 are illustrated in figure 1.8. Case A is the only one with complete information. For all the other cases, either E_1, E_2 or both are outside the interval $[\tau_0, \tau_1]$. Note that, in cases D, F and G the time at which event E_1 occurred is known even though it is in the pre-observation period (e.g., E_1 is assessed retrospectively). Cases B and C both illustrate right-censored observations; that is, E_1 is observed but E_2 is only known to be greater than a certain value. Right-censoring may arise from the termination of the observation window (case B) or the loss to follow-up (case C). Left-truncation is illustrated by cases D and E where both individuals have been at risk of an E_2 event for some time period before the start of the observation window. The distinction between the two is that for case D the time of event E_1 is known while it is unknown for case E. This difference between cases D and E greatly influences the statistical methods to use for inference; see section 1.4.2. In cases F and G, both E_1 and E_2 are not in the observation window. Case F is an example of a sojourn that is both left-truncated (with time of event E_1 known, in the present cases) and right-censored; this type of sojourn is often said to be doubly-censored (e.g., Kalton et al. (1992)). In case G, the time of event E_1 is known and the time of event E_2 is only known to be smaller than τ_0, this sojourn is thus said to be left-censored; see section 1.4 for further explanations on left-censoring and how it differs from left-truncation. Finally, case H illustrates a sojourn that does not get observed or recorded at all.

The results and theorems given in this thesis are derived under the assumptions of independent and noninformative right-censoring. Moreover, all right-censoring times C_i or τ_{i1} are stopping times; see section 1.3.2. Hence, censoring may depend only on past and not on future events. The independent censoring assumption specifies that the

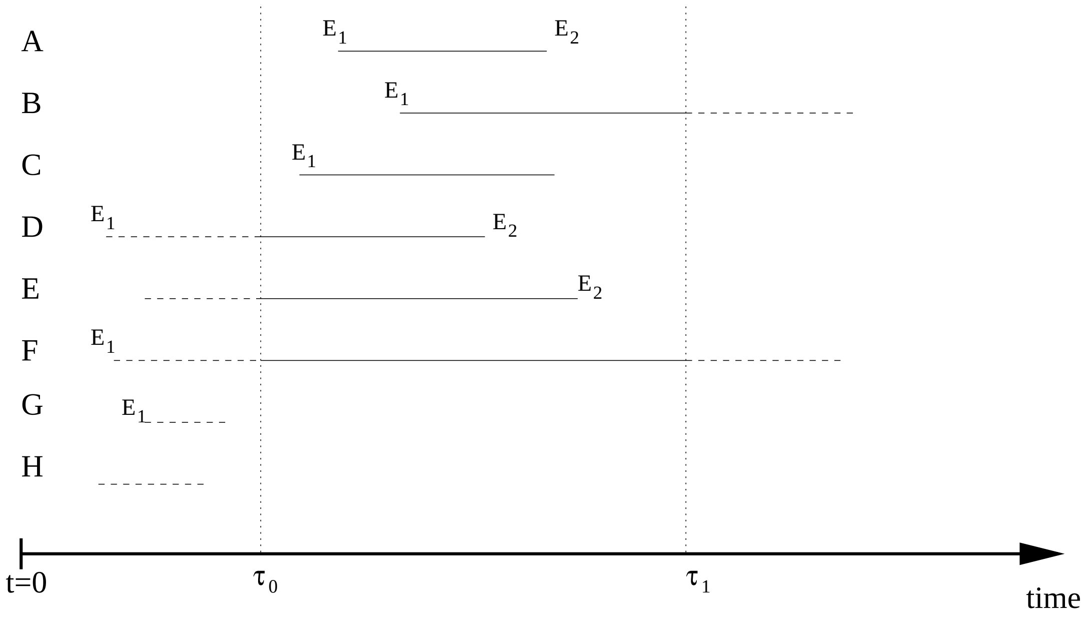

Figure 1.8: Scenarios of complete and incomplete observations.

probability of a transition in the next small time interval Δt, given no transition up to time t, is the same for individuals in general and those whose censoring time is greater than t; that is,

$$\Pr\{t \leq T < t + \Delta t \mid T \geq t\} = \Pr\{t \leq T < t + \Delta t \mid T \geq t, C \geq t\} \ , \qquad (1.13)$$

where T is the random variable corresponding to the time since the origin of the process (or the transition time for semi-Markov models) and C is the random variable corresponding to the right-censoring time. Independent censoring is also referred to as random censoring by some authors (e.g., Cox & Oakes (1984) and Klein & Moeschberger (1997)). Note that (1.13) is weaker than requiring independence between transition and censoring times. Clearly, if they are independent random variables, then the independent right-censoring assumption is satisfied. Noninformative right-censoring implies that the distribution of the C_i's does not depend on the parameters of the failure time distribution of the T_i's ($i = 1, \ldots, n$); see Andersen et al. (1993), section III.2.3, or Kalbfleisch & Prentice (2002), page 195.

It is important to mention that the time scale on which the different events are recorded is assumed known or given. It can be calendar time or something specific to the individual, such as age, time since a person got married or time since a disease was diagnosed.

In the last two decades, counting process and martingale theories have revolutionized event history and survival analyses; allowing for numerous new developments. As the methods proposed in chapter 3 are based on these theories, they will be the subject of the next section.

1.3 Basis of martingale framework

The counting process approach to survival and event history analyses was pioneered by Odd Aalen in the 1970's. This section introduces the notions of filtration, counting processes, stopping times, martingales, compensators, local martingales, predictable and optional variation processes, martingale central limit theorem and stochastic integration. The methods proposed in chapter 3 are based on semi-Markov models, as described in section 1.2.3, and assume that only one event or transition is possible or of interest; that is, $J = 1$. Hence, $N_{ij}(t)$, given by (1.2), simplifies to $N_i(t)$ and can be re-expressed to explicitly include right-censoring; that is,

$$N_i(t) = I(T_i \leq t, \delta_i = 1) \ , \tag{1.14}$$

where T_i is the random variable corresponding to the time elapsed since the last transition and

$$\delta_i = \begin{cases} 1 & \text{if individual } i \text{ experiences the end event} \\ 0 & \text{if individual } i \text{ is right-censored .} \end{cases}$$

Similarly, $\lambda_{ij}(t \mid \mathrm{H}_i(t^-))$, given by (1.4), simplifies to $\lambda_i(t)$, which is (1.9) without covariates. Since the martingale framework plays an important role in chapter 3, the notions of section 1.3 are established in terms of the above setting. See Andersen et al. (1993), section II.4, for a discussion of multivariate counting processes (i.e., $J > 1$).

1.3.1 Counting processes

A counting process is a stochastic process which keeps track of the numbers of occurrences of disjoint discrete events over time. Let $(\Omega, \mathcal{F}, P)$ be a probability space, then a family of σ-algebras of $\mathcal{F}$ is called a *filtration* and denoted by $\{\mathcal{F}_t; t \geq 0\}$, if it satisfies *les conditions habitulles*:

$$\begin{aligned} &1. \quad \mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F} \quad \forall\, s < t \\ &2. \quad \mathcal{F}_s = \bigcap_{t>s} \mathcal{F}_t \quad \forall\, s \qquad (1.15) \\ &3. \quad \text{let } A \subset B \in \mathcal{F}\,;\text{ then, } \Pr(B) = 0 \Rightarrow A \in \mathcal{F}_0\,. \end{aligned}$$

The first condition implies that $\{\mathcal{F}_t; t \geq 0\}$ is increasing, the second that it is right continuous, and the third that it is complete (i.e., for every t, $\mathcal{F}_t$ contains all P-null sets of $\mathcal{F}$). However, the third assumption can be omitted; see Andersen et al. (1993), page 117. The σ-algebras of $\mathcal{F}_t$ contains all events, including the null set, whose occurrence or not is determined by time t. More important is the pre-t σ-algebras $\mathcal{F}_{t-}$, which is the smallest σ-algebra containing all $\mathcal{F}_s$, with $s < t$. In other words, $\mathcal{F}_{t-}$ contains all events fixed strictly before time t, which is the same as the history $\mathrm{H}(t^-)$ defined on page 15. Therefore, history and filtration have the same meaning.

A *stochastic process* $\boldsymbol{Z} = \{Z(t); t \geq 0\}$ is a collection of time indexed random variables. It is said *adapted* to the filtration $\{\mathcal{F}_t; t \geq 0\}$ if $Z(t)$ is $\mathcal{F}_t$ measurable for all t's. The random functions $Z(\cdot, \omega) : \Re^+ \to \Re$, with $\omega \in \Omega$, are called the *sample paths* of Z, and $Z(t, \omega)$ can also be viewed as a function of t for fixed ω. The stochastic process $\boldsymbol{Z}$ is called *cadlag* (*continu à droite, limité à gauche*) if its sample paths are right-continuous with left-hand limits for almost all ω.

Definition 1.1 (Counting process) Let $\{\mathcal{F}_t; t \geq 0\}$ be a filtration on a probability space $(\Omega, \mathcal{F}, P)$ and assume that this filtration satisfies conditions (1.15), except maybe completeness. A stochastic process $\boldsymbol{N} = \{N(t); t \geq 0\}$ is called a *counting process* if it is an adapted cadlag process such that $N(0) = 0$ and $N(t) < \infty$ a.s., and that its sample paths are piecewise constant, nondecreasing and have jumps of size +1 only. □

Let $\mathcal{N}_t = \sigma(N_i(s), \gamma_i(s); i = 1, \ldots, n\,, 0 \leq s \leq t)$, where $\gamma_i(t) = I(t_i \geq t)$ is the indicator if individual i is still under study. Definition 1.1 implies that $\boldsymbol{N}_i = \{N_i(t); t \geq 0\}$, where $N_i(t)$ was defined by (1.14), are counting processes for $i = 1, \ldots, n$, with respect to the filtration generated by $\mathcal{F}_t = \mathcal{N}_t \vee \mathcal{F}_0$. That filtration is automatically right-continuous, increasing and can be made complete if necessary or desired, see (1.15). It also implies that $\boldsymbol{N}_\bullet = \{N_\bullet(t); t \geq 0\}$, where $N_\bullet(t) = \sum_{i=1}^n N_i(t)$, is a counting process. Note that $E(dN_i(t) \mid \mathcal{F}_{t^-}) = \gamma_i(t)\lambda_i(t)\,dt$.

1.3.2 Stopping times

A stopping time τ is a nonnegative r.v. (possibly degenerated) if, for each t, it can be established whether or not τ has occurred by time t knowing only the information in $\mathcal{F}_t$. In other words, thinking of τ as the time an event occurs, the decision to stop at time t is based only on the values of the process up to time t. This concept of stopping time is formalized in definition 1.2.

Definition 1.2 (Stopping time) Let $\{\mathcal{F}_t; t \geq 0\}$ be a filtration on a probability space $(\Omega, \mathcal{F}, P)$ and assume that this filtration satisfies conditions (1.15), except maybe completeness. A *stopping time* τ is a nonnegative r.v. taking values in $\Re^+$ such that $\{\omega : \tau(\omega) \leq t\} \in \mathcal{F}_t$ for all $t \geq 0$. □

Note that any fixed time τ is a stopping time. In addition to their obvious role in censoring, stopping times are important through the notion of localization. Many processes have some properties only up to a stopping time. The idea is to derive the result for the stopped process and extend it to the original process by letting the stopping times increase to an arbitrarily large value. First, definitions of a stopped process and of a localizing sequence of stopping times are given. Let $\boldsymbol{Z}$ be a stochastic process and τ be a stopping time, the *stopped process* $\boldsymbol{Z}^\tau$ is defined by $Z^\tau(t) = Z(t \wedge \tau)$, where $t \wedge s = \min(t, s)$. Note that if $\boldsymbol{Z}$ is cadlag and adapted, so is $\boldsymbol{Z}^\tau$. A nondecreasing sequence of stopping times τ_n is said to be a *localizing sequence* of stopping times if it satisfies

$$\Pr\{\tau_n \geq t\} \to 1 \quad \text{as } n \to \infty \text{ for all } t \in \Re^+ \ . \tag{1.16}$$

1.3.3 Local martingales

Martingales arise when a counting process $\boldsymbol{N}$ is expressed as the sum of a predictable systematic process and a zero-mean random noise. In other words, from writing

$$M(t) = N(t) - \Lambda(t) \ , \tag{1.17}$$

where $\boldsymbol{M} = \{M(t); t \geq 0\}$ is called the *counting process martingale*, and $\Lambda(t)$ is called the *cumulative intensity process* or *compensator* and is defined by (1.19). Note that a consequence of (1.17) is that $E(dM(t) \mid \mathcal{F}_{t^-}) = 0$ for all t, which is a characteristic of martingales.

Many results and applications of martingale theory require certain integrability conditions (e.g, (1.17) requires that $E(N(t)) < \infty \ \ \forall\, t$'s). These conditions may be difficult to verify in practice (e.g., when proving that the compensator of $\boldsymbol{N}$ and $\boldsymbol{M}^2$ are identical) or they might not even be true in certain cases. Localization allows for these conditions to be relaxed and produce results (e.g., limit theory) that are valid in a more general setting. Combining the notions of stopped processes and of localizing sequences yields definition 1.3.

Definition 1.3 (Local martingale) Let $\{\mathcal{F}_t; t \geq 0\}$ be a filtration on a probability space $(\Omega, \mathcal{F}, P)$ and assume that this filtration satisfies conditions (1.15), except maybe completeness. A *local martingale* is an adapted cadlag stochastic process $\boldsymbol{M} = \{M(t); t \geq 0\}$ such that there exist an increasing localizing sequence of stopping times τ_n, as defined by (1.16). This sequence is such that the stopped processes $I(\tau_n > 0)\boldsymbol{M}^{\tau_n}$ are martingales for each n. □

Note that any local martingale is locally uniformly integrable. Similarly, a *local square integrable martingale* is defined as a stochastic process $\boldsymbol{M}$, as described in definition 1.3, such that the stopped processes $I(\tau_n > 0)\boldsymbol{M}^{\tau_n}$ are square integrable martingales for each n; that is,

$$\sup_{t \in \Re^+} E\Big(\big(I(\tau_n > 0)\boldsymbol{M}^{\tau_n}(t)\big)^2\Big) < \infty \quad \text{for each } n \; . \tag{1.18}$$

Putting together the notions of sections 1.3.1 and 1.3.2 yields the following momentous result. Since $\boldsymbol{N}_i$ is a counting process, it is adapted, cadlag and nondecreasing for $i = 1, \ldots, n$. If it is further assumed that $\boldsymbol{N}_i$ is locally bounded[3] (i.e., $0 \leq \boldsymbol{N}_i^{\tau_n} \leq n$), then it is a local submartingale and has a nondecreasing predictable process $\boldsymbol{\Lambda}_i = \{\Lambda_i(t); t \geq 0\}$, where

$$\Lambda_i(t) = \int_0^t \gamma_i(u)\, \lambda_i(u)\, du \; . \tag{1.19}$$

This predictable process is such that

$$\boldsymbol{M}_i = \boldsymbol{N}_i - \boldsymbol{\Lambda}_i \tag{1.20}$$

is a *local counting process martingale.* Moreover, it can be shown that $\boldsymbol{M}_i$ is a local square integrable martingale. This result, in particular (1.20), is a consequence of the extended Doob-Meyer decomposition theorem; see Fleming & Harrington (1991), page 58, for further details. Note that (1.20) implies that $\boldsymbol{M}_\bullet = \boldsymbol{N}_\bullet - \boldsymbol{\Lambda}_\bullet$ is also a local square integrable martingale, where $M_\bullet(t) = \sum_{i=1}^n M_i(t)$ and $\Lambda_\bullet(t)$ is its corresponding compensator.

As they have many very useful properties, martingales and local martingales play a key role in the general theory of stochastic processes. As in (1.17) and (1.20), they allow splitting counting processes into a systematic and a random parts. This decomposition is based on the he Doob-Meyer decomposition theorem (e.g., Fleming & Harrington (1991), page 37), or its extended version. They also satisfy the martingale property

$$E(M(t) \mid \mathcal{F}_s) = M(s) \quad \text{for all } s \leq t \; , \tag{1.21}$$

which is at the core of their definitions. Local counting process martingales, as defined

[3] See Andersen et al. (1993), page 73, for further explanations on the construction of the localizing sequence $0 < \tau_1 \leq \tau_2 \leq \tau_3 \leq \ldots$, which is based on the notion of marked point process.

by (1.20), have further useful properties. In particular, when $\boldsymbol{\Lambda}_i$ is continuous,

$$<\boldsymbol{M}_i> = \boldsymbol{\Lambda}_i \tag{1.22}$$

$$<\boldsymbol{M}_i, \boldsymbol{M}_{i'}> = 0 \quad \text{for } i \neq i' \ , \tag{1.23}$$

where $<\boldsymbol{M}>$ is called the *predictable variation process* of $\boldsymbol{M}$ and is defined, in terms of differential notation, by $d<\boldsymbol{M}>(t) = \mathrm{Var}(dM(t) \mid \mathcal{F}_{t^-})$. Similarly, $<\boldsymbol{M}, \boldsymbol{M}'>$ is the *predictable covariation process* of $\boldsymbol{M}$ and $\boldsymbol{M}'$. This process is equal to zero, since the $N_i(t)$'s have no jump in common. Another process is the so called square brackets process or *optional variation process* $[\boldsymbol{M}]$ of $\boldsymbol{M}$, where $[M](t) = \sum_{s \leq t} \Delta M(s)^2 = \sum_{s \leq t}(M(s) - M(s^-))^2$. See Andersen et al. (1993), section II.3.2, for further details on predictable and optional processes. These properties imply, from (1.22), that $\boldsymbol{M}$ and $\boldsymbol{M}^2$ have the same compensator $\boldsymbol{\Lambda}$ and, from (1.23), that counting process martingales are orthogonal.

1.3.4 Martingale central limit theorem

As for the usual central limit theorem, there exists many versions; see Fleming & Harrington (1991), chapter 5. The one summarized by theorem 1.1 is due to Rebolledo (1980).

Theorem 1.1 (Rebolledo's martingale central limit theorem) For $n = 1, 2, \ldots$, let $\boldsymbol{M}^n$ be local square integrable martingales defined on $(\Omega^n, \mathcal{F}^n, P)$ with respect to filtration $\{\mathcal{F}^n_t; t \geq 0\}$. For each $\varepsilon > 0$, let $\boldsymbol{M}^n_\varepsilon$ be square local martingales, containing all the jumps of $\boldsymbol{M}^n$ larger than ε in absolute value. Furthermore, let $\boldsymbol{M}^{\boldsymbol{\infty}}$ be a continuous Gaussian martingale with mean zero and covariance function $A(\min(s,t))$, where $\boldsymbol{A} =< \boldsymbol{M}^\infty >= [\boldsymbol{M}^\infty]$. Given $\boldsymbol{A}$, it can be shown that the process $\boldsymbol{M}^{\boldsymbol{\infty}}$ always exists. Consider the following conditions

$$<M^n> (t) \xrightarrow{P} A(t) \quad \forall\, t \in \Re^+ \text{ as } n \to \infty \tag{1.24}$$

$$[M^n]\,(t) \xrightarrow{P} A(t) \quad \forall\, t \in \Re^+ \text{ as } n \to \infty \tag{1.25}$$

$$<M^n_\varepsilon> (t) \xrightarrow{P} 0 \quad \forall\, t \in \Re^+ \text{ and } \varepsilon > 0 \text{ as } n \to \infty \tag{1.26}$$

Then either (1.24) or (1.25) together with (1.26) imply

$$\boldsymbol{M}^n \xrightarrow{\mathcal{D}} \boldsymbol{M}^\infty \,; \tag{1.27}$$

moreover, both (1.24) and (1.25) then hold. □

1.3.5 Stochastic integration

Stochastic integration refers to the integral of one stochastic process with respect to another. For example, consider the following two stochastic processes $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$; then, $\int \boldsymbol{Z}_1\, d\boldsymbol{Z}_2$ denotes the new stochastic process

$$Z_3(t) = \int_0^t Z_1(u)\, dZ_2(u)\ , \tag{1.28}$$

defined for each $\omega \in \Omega$ and $t \in \Re^+$ such that

$$\int_0^t |Z_1(u,\omega)|\, |dZ_2(u,\omega)| < \infty\ .$$

If $\boldsymbol{Z}_1$ is a predictable process and $\boldsymbol{Z}_2$ a local martingale, then the stochastic process $\boldsymbol{Z}_3$, given by (1.28), has special properties as described in theorem 1.2.

Theorem 1.2 Let $\boldsymbol{M}$ be a local martingale and $\boldsymbol{H}$ a locally bounded predictable process; then,

$$\int \boldsymbol{H}\, d\boldsymbol{M}$$

is a local martingale and

$$\left[\int \boldsymbol{H}\, d\boldsymbol{M}\right] = \int \boldsymbol{H}^2\, d[\boldsymbol{M}]\ .$$

If it is further assumed that $\boldsymbol{M}$ is square integrable; then, the previous results hold and, in addition

$$\left\langle \int \boldsymbol{H}\, d\boldsymbol{M} \right\rangle = \int \boldsymbol{H}^2\, d{<}\boldsymbol{M}{>} \quad .$$

□

Andersen et al. (1993), chapter II, contains an excellent discussion of counting processes, martingales and stochastic integration. Further explanations can be found in Fleming & Harrington (1991), chapters 1, 2 and 5. Methods for handling left-truncated observations are proposed and compared in chapter 5. In the next section, key concepts of left-truncation are introduced and exemplified.

1.4 Basis of left-truncation

Left-truncation is another source of incomplete information and is portrayed in figure 1.9, where the time of the 1 → 2 transition is left-truncated. An observation is said to be left-truncated if it has been at risk of experiencing an event or transition for some time period before coming under study and that event or transition has not occurred prior to the start of the study. The time period in question may or may not be known. In either case, had the observation experienced the event in question before the beginning of the observation window, it would not have been sampled. Left-truncated data are incomplete because they do not include observations that have not survived long enough to be observed. Hence, they tend to over-represent low risk cases, leading to sample selection bias. This is a crucial and important difference between truncation and censoring. A good summary of the problems caused by left-truncated data and possible solutions is given by Guo (1993). After presenting characteristics and challenges of left-truncated data, he discusses statistical methods for both the cases of unknown and known start times.

Left-truncation arises in many different settings. Social science provides numerous examples. Suppose the interest is in studying time until marriage dissolution; namely,

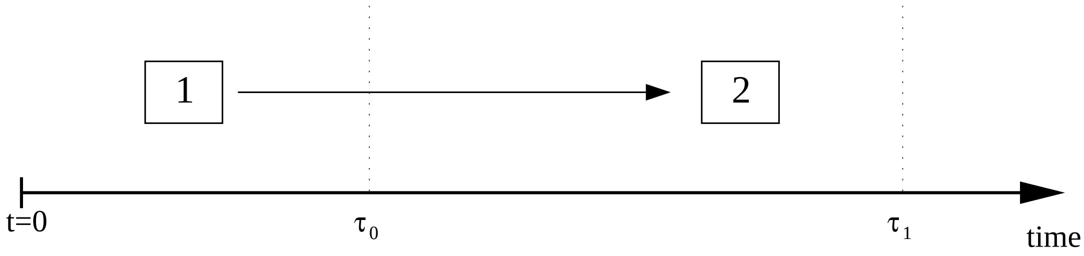

Figure 1.9: Left-truncation.

time between the wedding day (starting point) and the divorce settlement date or death (end event). To this end, a random sample of individuals is selected and followed over some time period. Unless the sample is restricted to single or young people, many sampled individuals were married before the start of the study. These individuals were at risk of "failing in their marriage" before becoming under study and tend to over-represent low divorce rate cases. Marriage dates can easily be found retrospectively, thus the time period prior to the start of the observation window is known in the marriage dissolution example. However, in modeling the time from HIV infection to the onset of AIDS it is very difficult to determine when an individual got infected by the HIV virus. Hence, there are many left-truncated observations with unknown exposure time prior to the observation window. Other examples of left-truncation are given by Turnbull (1976), Hyde (1977, 1980) and Andersen et al. (1982).

In survival analysis the term left-truncation applies to an individual or a subject. As event history analysis deals with more than one event or transition, the term left-truncation typically applies to a sojourn experienced by an individual and not to the entire history process of that individual. A *sojourn* is defined as the time spent in a given state (multi-state framework) or the time between two events (event occurrence framework).

Left-censoring and left-truncation are distinct types of incomplete data and different statistical methods are used to handle left-censored and left-truncated observations; thus, it is important to distinguish between the two. Although less common, incomplete data can also arise from left-censoring. This was illustrated by case G in figure 1.8. Turnbull (1974) gives a good example of left-censoring, which involves data from a study by Leiderman et al. (1973) on infant precocity. The goal was to create norms for infant development for a community in Kenya with the objective of comparing these to United States and United Kingdom standards. Starting in January 1970, each child in a sample of 65 children (born between July 1 and December 31 of the previous year) was given a monthly test to see if he/she was able to perform a given task. Left-censoring occurs when a child is found to be able to perform the task at the first test. Unlike left-truncated observations, left-censored observations remain in the sample even though they have experienced the end event, completion of a given task in this case. In summary, truncation consists of "sampling from an incomplete population"; that is, from a conditional distribution (conditional on the fact that failure has not occurred prior to the time of selection τ_0). On the other hand, left-censoring arises when observations are sampled from a complete population, but duration or failure times below a given value are left unspecified.

Hald (1952), page 144, was the first to systematically distinguish between truncated and censored data. Left-truncation has also been referred to as delayed entry (e.g, Keiding (1992)), initial censoring (e.g, Kalton et al. (1992)), partial left censoring (e.g, Tuma & Hannan (1984) and Yamaguchi (1991)) and initial conditions (e.g, Heckman & Singer (1986)). Note that, from now on, we will only use the term left-truncation. Moreover, left-censoring will not be discussed further in this thesis.

1.4.1 Independent left-truncation

For the equations of section 1.4.2 to hold, in addition to the assumption of independent right-censoring, the left-truncation mechanism must also be independent. This means that the hazard of left-truncated people is the same as the rest of the population. More formally,

$$\Pr\{t \leq T < t + \Delta t \mid T \geq t, L, T > L\} = \Pr\{t \leq T < t + \Delta t \mid T \geq t\} \tag{1.29}$$

where T is the random variable corresponding to the transition time and L is the random variable corresponding to the left-truncation time (or entry time). Deviation from the independent left-truncation assumption may occur when truncation and transition times are related to each other. A formal test for independence of truncation and survival times, based on Kendall's τ test, was introduced by Tsai (1990). See Kalbfleisch & Lawless (1991) for further discussion on the implications of (1.29).

1.4.2 Inference with left-truncation

This section summarizes three categories of methods for conducting statistical inference with left-truncated sojourns, which will be studied and compared in chapter 5. These methods require different amounts of information about the time prior to the start of the observation window; see section 1.2.4. They are derived under the assumptions of independent right-censoring and independent left-truncation. For simplicity, the following discussion is expressed in terms of survival analysis models; see section 1.2.2. However, the concepts described in section 1.4.2 are valid for a wide range of event history models (e.g., semi-Markov or sojourn duration models of section 1.2.3). These methods account for the selection bias created by sampling from an incomplete population or from a conditional distribution, where some individuals were at risk of experiencing the end event prior to τ_{i0}. Failure to recognize and take into account this feature results in a biased likelihood function, where low risk transitions are over-represented.

The methods of the first category are based on a conditional likelihood approach, which has been around for many years; see Schoen (1975), Thompson (1977) and Hyde (1977, 1980) for early applications. Let $t_1, \ldots, t_n$ be left-truncated and possibly right-censored observations. These observations having been at risk for $b_1, \ldots, b_n$ time units prior to the start of the observation windows $[\tau_{i0}, \tau_{i1}]$; thus, the conditional approach. This situation is portrayed in figure 1.10.

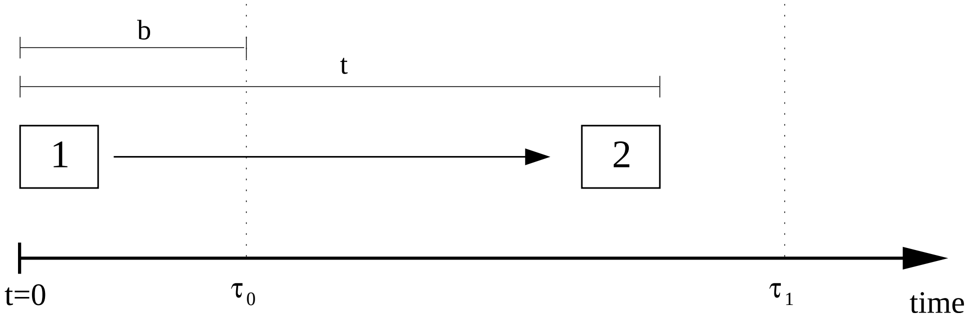

Figure 1.10: Conditional approach for known exposure time.

Let τ_{i0} and τ_{i1} be stopping times with respect to the history processes $\{Y_i(t); t \geq 0\}$, for $i = 1, \ldots, n$. For example, an individual may be lost to follow-up at some random time τ_{i1} after the start of the study. Similarly, an individual may join an existing household after being exposed to a given risk for some random time period τ_{i0}. Then, a partial likelihood on which inference about the duration or lifetime distribution can be based is

$$L = \prod_{i=1}^{n} \lambda_i(t_i)^{\delta_i} \frac{S_i(t_i)}{S_i(b_i)} \ , \tag{1.30}$$

where $\lambda_i(t)$, $S_i(t)$ and δ_i were, respectively, defined in (1.9), (1.10) and (1.14). To simplify notation, covariates were left out of (1.30)–(1.32) and section 1.4.3.

The methods of the second category further assume that the b_i's are observations from i.i.d. random variables B_i's, which correspond to the backward recurrence or left-truncation times. Their p.d.f. is $f_{B_i}(b) = S_i(b)/\mu$, where μ is the mean of the "untruncated" random lifetime distribution corresponding to T_i. Incorporating that additional information, (1.30) becomes

$$L = \prod_{i=1}^{n} f_{B_i}(b_i)\, \lambda_i(t_i)^{\delta_i} \frac{S_i(t_i)}{S_i(b_i)} \ . \tag{1.31}$$

The third category of statistical methods assumes that processes are in equilibrium. Let $\tilde{t}_1, \ldots, \tilde{t}_n$ be observations, possibly right-censored, from the forward recurrence time distribution corresponding to T_i; that is, $t_i = b_i + \tilde{t}_i$ for $i = 1, \ldots, n$; see figures 1.10 and 1.11. The p.d.f. of $\widetilde{T}_i$ is $f_{\widetilde{T}_i}(t) = S_i(t)/\mu$. For these left-truncated observations, it is only known how long they remained alive after the start of their respective observation windows $[\tau_{i0}, \tau_{i1}]$; b_i and thus the total length of their respective durations are unknown. This is portrayed in figure 1.11. Under the equilibrium assumption, a likelihood on which inference can be based is

$$L = \prod_{i=1}^{n} \left(\frac{1}{\mu} S_i(\tilde{t}_i)\right)^{\delta_i} \left(\int_{\tilde{t}_i}^{\infty} \frac{u}{\mu} S_i(u)\, du\right)^{1-\delta_i} . \tag{1.32}$$

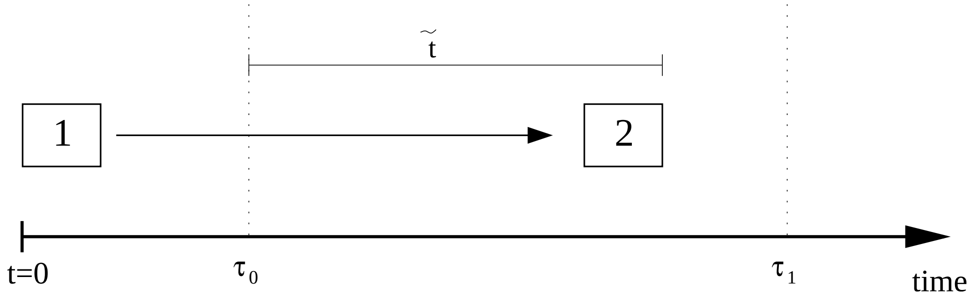

Figure 1.11: Equilibrium approach for unknown exposure time.

As mentioned previously, note that (1.30), (1.31) and (1.32) do not all need the same data. In particular, (1.30) and (1.31) need the values of the b_i's, but (1.32) does not. Also, note that (1.31) requires the additional assumption that the B_i's have a known density and (1.32) that the process is in equilibrium, whereas (1.30) does not. Finally, if the process is in equilibrium, $f_B(b)$ in (1.31) has the same p.d.f. as $\widetilde{T}_i$ and so it provides additional information about $S_i(t)$; see section 5.2 and, in particular, (5.4).

1.4.3 Left-truncated counting processes

This section formalizes the definition of left-truncation. Let $N_i(t)$, as defined by (1.14), be a counting process on a probability space $(\Omega, \mathcal{F}, P)$ with respect to the filtration $\{\mathcal{F}_t; t \geq 0\}$, where $\mathcal{F}_t = \mathcal{F}_0 \vee \mathcal{N}_t$. Then, $\boldsymbol{M}_i = \boldsymbol{N}_i - \boldsymbol{\Lambda}_i$, as defined by (1.20), is a local (square integrable) martingale.

The following definition is given under the assumptions of independent left-truncation (see section 1.4.1) and the existence of a larger filtration $\mathcal{G}_t \supseteq \mathcal{F}_t$, such that the compensator of $\boldsymbol{N}_i$ with respect to $\{\mathcal{G}_t; t \geq 0\}$ is also $\boldsymbol{\Lambda}_i$. This larger filtration includes all possible extra information about random variables involved in the truncation time. Let τ_{i0} be a stopping time with respect to this larger filtration and consider the event $A \in \mathcal{G}_\tau$ with $P(A) > 0$. Given that event A has actually occurred prior to time τ_{i0},

$$N_i^\tau(t) = N_i(t) - N_i(t \wedge \tau_{i0}) \ , \qquad (1.33)$$

where $t \wedge s = \min(t, s)$, is called a *left-truncated counting process.* It keeps track of the number of events experienced by individual i in the interval $[\tau_{i0}, t]$ given the occurrence of event A. For example, A could be the entry into first marital union in the marriage dissolution example or the time of HIV infection in the AIDS example of page 33.

Note that (1.33) has intensity process $\lambda_i^\tau(t) = \lambda_i(t)\, I(t > \tau_{i0})$ with respect to the filtration $\mathcal{G}_t^\tau = \mathcal{G}_t \vee \mathcal{G}_\tau$. Keiding (1992) provides an insightful discussion and gives examples on independent left-truncation. Andersen et al. (1993), section III.3, also give further explanations on the subject.

Finally, it is possible to discard left-truncated sojourns or observations from the sample, when τ_{i0} is a stopping time; see Aalen & Husebye (1991). This may result in a substantial loss of information in some settings, but avoids the need for additional assumptions when the entry times or the b_i's (see figure 1.10) are unknown.

Chapter 2

Particularities of longitudinal surveys

This chapter discusses particularities and special features of longitudinal surveys. The increase in the availability of administrative data has made such surveys easier to produce and more cost effective. They have become more appealing to governments and other organizations or people interested in understanding various patterns of social, health and economic changes over time. A considerable amount of event history data is collected through longitudinal studies which involve the use of complex survey designs such as clustering and stratification. This use of complex survey designs, the dynamic nature of longitudinal surveys and the numerous goals or objectives they have to meet, complicate their analyses.

Section 2.1 describes and illustrates seam-effect problems caused by the use of retrospective information. Section 2.2 exemplifies the prevalence of left-truncated observations through SIPP and PSID data. Section 2.3 discusses attrition and non-response. Section 2.4 explains some problems with top-up samples, a common technique used to “counter-balance” the effects of attrition. Section 2.5 talks about the dynamic nature of longitudinal surveys. Section 2.6 is devoted to the complex and controversial topic of defining proper sampling weights. As mentioned in chapter 1, event history analysis deals with the modeling and the analysis of the history processes of individuals and, even though longitudinal studies involve complex survey designs, the goal of such studies is often analytical rather than descriptive. Analytical inference based on longitudinal survey data is discussed in section 2.7.

The above only covers some of the particularities of longitudinal surveys. Further complications include: intermittent follow-up of individuals, right-censoring, non-ignorable censoring, missing or miss-measured observations, respondents burden, record linkage, non-response that does not occur completely at random, frequent use of proxy information, confidentiality issues, etc. Binder (1998) contains a good overview of the particularities longitudinal surveys. The authors in Kasprzyk et al. (1989) also provide interesting and, sometimes, controversial discussions on these topics.

2.1 Seam-effect

The seam-effect is a form of recall bias or non-sampling error unique to longitudinal surveys. It refers to a disproportionately high occurrence of transitions at the "seam" between two reference periods, generally referred to as waves. In other words, the number of participant-reported transitions is much greater between dates for which the data were collected in different waves than between dates for which the data were collected in the same wave. In addition to misplacing the beginning or end of a spell, some participants might completely forget to report a spell. The seam-effect will likely bias the estimation of spell durations, as end and start dates are not reported properly.

The extent and consequences of the seam-effect are best illustrated using the 1987 SIPP data. In section 1.1.2, it was mentioned that interviews were being conducted at four month intervals. Evidence of the seam-effect is the considerably larger numbers of reported transitions in and out of the different social programs between months 4 and 5, 8 and 9, 12 and 13, etc. than between months 1 and 2, 2 and 3, 3 and 4, 5 and 6, etc. In addition, if there was no seam-effect, the proportions of spell starts and stops would be around 25% for each month of recall. However, these proportions vary from 3.5% to 92.9%; see Kalton et al. (1992), tables 1.4 and 1.5. From these tables, it can also be seen that for the 4th month of recall these proportions are over 50% for most social transfer programs and spells without health insurance, another clear indication of the seam-effect. The same phenomenon can also be observed by looking at figure 2.1; see section 4.1.2 for further discussion on the duration of spells without health insurance coverage. The seam-effect is shown by the sharp decreases in the Kaplan-Meier curve before the various seams represented by dotted lines. Note that these are more important for shorter spells (i.e., one year or less). This is mainly due to the fact that sample sizes are larger in these cases. The above is in agreement with the results of Kalton et al. (1992), table 1.15 for spells without health insurance and tables 1.6–1.14 for spells on social transfer programs. For further discussions on the seam-effect in SIPP see Kalton & Miller (1991).

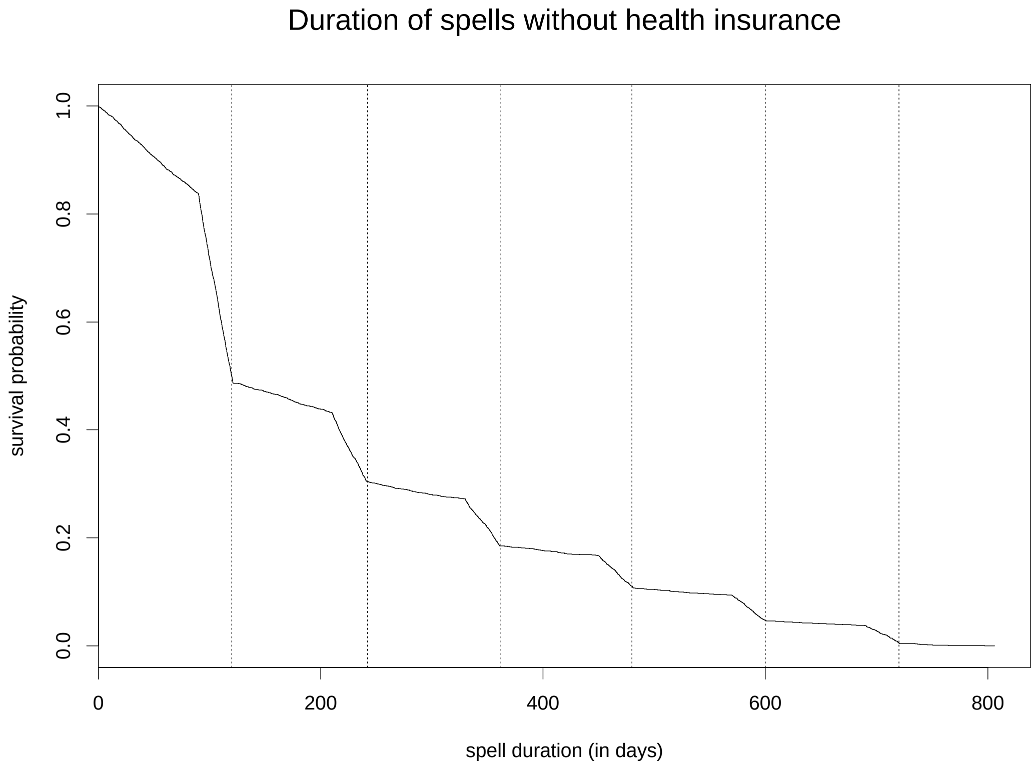

Figure 2.1: Seam-effect in the 1987 SIPP data.

There are few methods for either preventing or adjusting seam-effect bias. Prevention is mainly done through the use of computer assisted interviews (CAI), which is also referred to as dependent interviewing by some authors. Lemaître's (1992) comparisons of two feedback techniques yielded that computer assisted interviews is the only reliable method for collecting consistent data over time. The effectiveness of CAI is exemplified by the very high number of job starts in January 1994 (i.e., 800,000 compared to 300,000 in January 1993 or slightly over 200,000 in January 1995), when about 30% of the SLID sample could not be interviewed with the help of CAI due to computer difficulties. See Cotton & Giles (1998), section 2.2, for further explanations on what happened in January 1994. Since CAI is such an effective method for reducing the seam-effect, it is not surprising that it is used in SLID and, starting with the 1996 panel, in SIPP. In particular, the seam-effect in SLID is greatly diminished by the use of CAI; see figure 4.17 and Cotton & Giles (1998). Unfortunately, this was not the case for the 1987 SIPP data and Kalton et al. (1992) had to adjust for the seam-effect. Their method, which is based on Young's (1989) constant wave response idea, is described in section 2.2 of their technical report. It consists in reallocating a proportion of the starts and stops reported at 4 months of recall. For each participant reporting a transition at the seam, its sampling weight was divided into 4 parts, one for each month of recall. This was done by estimating the probabilities that the actual month of recall was i given it was reported as month 4, for $i = 1, 2, 3$ and 4. Even though this method does not account for response errors off seam and for short spells of program participation, it yields good results; see Kalton et al. (1992), tables 2.3–2.12.

2.2 Left-truncation

Left-truncation was introduced in section 1.4 and will be discussed further in chapter 5. This section focuses on the prevalence of left-truncated sojourns in SIPP (see section 1.1.2) and in PSID (see section 1.1.3). Table 2.1 shows the numbers and percentages of left-truncated sojourns and of sojourns that are both left-truncated and right-censored in the 1987 SIPP panel. Discarding those sojourns, as suggested by some authors (e.g., Allison (1984)), would result in a considerable reduction of sample sizes. Kalton et al. (1992), table 1.3 and chapter 4, give further information on left-truncation in the 1987 SIPP panel and interesting methods for including these truncated sojourns.

	truncated only†		truncated & censored‡	
Program	number	percentage	number	percentage
General assistance	98	28.4%	54	15.7%
Food stamps	930	26.6%	803	22.9%
Social Security	278	5.8%	3,563	74.2%
State unemployment	241	13.9%	0	0.0%
No health insurance	2,743	25.3%	2,651	24.4%
Veterans compensation	128	17.9%	281	39.2%

† Sojourns that are left-truncated, but not right-censored;

‡ Sojourns that are both left-truncated and right-censored.

Table 2.1: Prevalence of left-truncation in the 1987 SIPP data.

The marital dissolution data of PSID is another example of the prevalence of left-truncation. In that dataset, 49% of white women and 39% of black women were in their first marital union prior to 1968, the start of the observation window. In addition to sample size reductions, it could be argued that by discarding left-truncated sojourns information on individuals presenting unique characteristics may be lost. In the first marital union duration example, it is easy to retrospectively find the dates on which these left-truncated women got married and derive the durations of the corresponding unions. These provide some unique answers concerning divorce rates for marital unions that lasted longer than 30 years.

2.3 Attrition

Attrition is a form of non-response particular to longitudinal surveys. It refers to the shrinking of the sample or the loss of participants as time goes on. As noted by Groves (1989), non-response can be divided into various categories, depending on the reasons a response could not be obtained (e.g., refusal, lost to follow-up, unable to trace and other reasons). Distinguishing between these categories of non-response is important in longitudinal surveys. At the beginning, attrition is mainly due to refusals, but this tends to decrease in later waves. On the other hand, more participants are lost to follow-up as time goes on. Longitudinal surveys are fortunate since they have considerable information on the non-respondents who participated in earlier waves, which can be used to study and to model non-response. However, they suffer from two additional and unique forms of non-response called "wave non-response" and "partial household non-response". The former arises when a sampled individual responds for some but not all of the waves for which he/she was eligible. The latter arises when some but not all eligible members of a sampled household respond. This implies that some household characteristics (e.g., total household earnings) can not be computed.

The importance of attrition in SLID is illustrated in table 2.2 for the various waves of the 1$^{\text{st}}$ panel (which started on January 1, 1993) and the 2$^{\text{nd}}$ panel (which started on January 1, 1996). Note that some response rates for the 2$^{\text{nd}}$ panel are not yet available. These are referred to as longitudinal response rates in SLID and correspond to the percentages of original sampled individuals who responded in each wave. From table 2.2, 82.7% of the original sample of the 1$^{\text{st}}$ panel remained in the survey at the end of that panel on December 31, 1998. See Michaud & Webber (1994) for a more in depth discussion of attrition and non-response in the Survey of Labour and Income Dynamics.

Another example of attrition comes from the Panel Study of Income Dynamics survey (PSID), described in section 1.1.3. After an initial loss of about 14% of sample members between the first and second waves, the annual sample attrition rate was around

Wave	Panel 1	Panel 2
1	93.3%	89.8%
2	89.6%	87.4%
3	86.5%	86.5%
4	85.2%	†
5	83.2%	†
6	82.7%	†

† Not yet available.

Table 2.2: Attrition rates in SLID

2%. Consequently, after 14 years, only 60% of the original sample remained for the 1981 wave. As mentioned by Lillard (1989), section 2.1, this has the potential of introducing substantial bias as the remaining sample becomes less and less representative over time. This is one of the reasons why both SLID and SIPP have been going to great length to follow all originally selected members when their household moves or splits. Unfortunately, the same is not true for people that joined existing households after the start of the study. Therefore, the attrition rate for these individuals is much higher.

2.4 Top-up sample

To "counter-balance" the effects of attrition, top-up samples consisting of non-sampled or non-randomly selected individuals are added to the original sample over time. Generally, these consist of individuals who joined (e.g., through birth or marital union) the household of original sampled participants and who were not part of the original sample at the start of the study.

How to assign appropriate sampling weights to these top-up individuals is still controversial and open for study; see section 2.6 and Lillard (1989), section 2.2. Even though the same information is collected on these individuals than on the original sampled ones, the current procedure in SLID[1], SIPP and PSID is to give these individuals a longitudinal sampling weight of zero. Therefore, they are excluded from any weighted longitudinal analyses. The methods proposed in chapter 3 are based on uninformative sampling designs and unweighted procedures; thus, top-up samples are easily and naturally incorporated. This is one of the advantages of the proposed methods. For example, when applying them to the 1987 SIPP dataset to analyze the duration of spells on the food stamps program, it was possible to utilize a total of 1,816 spells in one of our analyses of section 4.1.1; that is, 1,095 spells experienced by members of the original sample plus 721 spells from top-up samples. The increase in sample size is much greater when studying spells without health insurance. In that case, one of our analyses of section 4.1.2 was based on a total of 7,656 spells without health care coverage. Only 56% or 4,314 of them were experienced by members of the original sample, the others were experienced by individuals that joined existing household. Similarly, when studying the duration of jobless spells from the 1993 SLID dataset, unweighted analyses of section 4.2.1 can be based on 16,682 spells while weighted ones are restricted to 14,978 spells. The importance of top-up samples in PSID has varied greatly since 1968. In 1990, 58.8% of the 20,535 participants were top-ups. See Lillard (1989), section 2.2, and Kalton & Brick (1995), section 1, for further information on top-up samples in PSID.

2.5 Dynamic nature

This section builds on sections 2.3 and 2.4, and will influence sections 2.6 and 2.7. A unique characteristic of longitudinal surveys is that the population keeps changing over time. Participants move in and out of the target population during the life of the study. Other changes include births, deaths, attrition and top-up samples composed of non-sampled individuals joining existing households. In addition, some households cease to

[1] In SLID, these top-up individuals are referred to as “co-habitants”; see section 1.1.1.

exist (e.g, divorce), new ones are created (e.g., children moving to live on their own) and others undergo major changes. Hence, one could view data collected through longitudinal studies as generated by a super-population model or process and, at any given time t, there is a finite population U_t from which a sample could be drawn. However, this finite population is ever-changing with time. As noted by Folsom et al. (1989), section 2.4, this complicates every statistical procedure, even simple ones such as descriptive statistics like the annual mean or median. It also makes defining proper sampling weights more difficult, as discussed in section 2.6.

For these reasons, defining the target population for a longitudinal survey is more complex and open to discussion than for a cross-sectional survey. Goldstein (1979), section 2.1, proposed four ways of defining the target population of a longitudinal survey; they are:

1. All individuals who were living in the specified geographic or target area at the beginning of the study, who remained eligible and in that area at each subsequent waves;

2. The above individuals together with all the individuals who moved or were born in that area after the beginning of the study;

3. All individuals who were living in the specified geographic or target area at the beginning of the study, regardless of where they lived at subsequent waves;

4. All individuals who were living in the specified geographic or target area at the beginning of the study together with all the individuals who moved or were born in that area after the beginning of the study.

The first definition makes follow-up problems much easier. However, bias might be introduced as people leaving the target population might have different characteristics than the ones staying. The last definition is the most appropriate to answer questions regarding changes in individuals over time. It is also the most general and the one preferred in large scale longitudinal surveys like SLIP, SIPP and PSID.

2.6 Defining weights

The dynamic nature of the population under study greatly complicates the computation of sampling weights. Depending on the procedures for assigning sampling weights it is possible to account for one or more of the following: non-response, attrition, top-up samples, people moving in and out of the target population, modification in the compositions of various households, etc.

Defining sampling weights as the reciprocal of the "true" probabilities of selection generally does not work in longitudinal households surveys. As mentioned by Ernst (1989) and Kalton & Brick (1995), it is operationally impossible to compute these probabilities. For example, in SLID and SIPP, a household is part of the third wave of interviews if and only if at least one household member was part of the original sample. Therefore, to compute the probability of this event is essential to determine the "true" probability of selection of that household. This would require having the probability of selection at the time the original sample was selected for every member of the household, including people that joined that household afterwards. Then, compute the probability that at least one of these members was selected at the beginning of the longitudinal survey and is still in the target population at the time of the third wave of interviews. In SLID, this would have to be repeated for each of the more than 15,000 households that existed at some time point during the life of the 1$^{\text{st}}$ panel. Fortunately, this does not imply that there is no unbiased weighted estimator; see Ernst (1989), section 3. For example, households' sampling weights can be defined as the reciprocal of the probabilities of selection at the time the original sample was selected. These weights can then be modified as time goes on to account for non-response and attrition. The problems caused by top-up samples can easily be solved by assigning weights of zero to newborns and other individuals joining existing households after the start of the study. This is similar to the procedures used in SLID and SIPP.

As it is convenient and in “theory” does not create bias, assigning longitudinal weights of zero to newborns and other individuals joining existing households after the start of the study is common. Therefore, these individuals are often excluded from weighted analyses. For example, Kalton et al. (1992) estimated the duration of spells in various transfer programs from the 1987 SIPP survey. In their report, individuals that were not part of the original sample drawn in 1987 were assigned a sampling weight of zero. In some cases, this resulted in about 50% of the observations being discarded from the data (see the different sample sizes of table 4.3). Fortunately, it is possible to include these top-up individuals by assigning them non-zero sampling weights. Ernst (1989), Folsom et al. (1989), sections 2.3 and 2.4, and Kalton & Brick (1995) proposed various procedures to do so. Although most of these procedures are intended for cross-sectional studies from longitudinal datasets, they can be adapted to create non-zero longitudinal sampling weights for top-up samples. However, the rules for following these individuals are not the same as for the members of the original sample, which has the potential to create serious non-response bias.

In addition to the problems created by top-up samples, the sampling weights must be adjusted for attrition and non-response. This includes the two forms of non-response unique to longitudinal surveys discussed in section 2.3. In summary, defining longitudinal sampling weights can easily become complicated and there are many different methods one may choose from. This leads to the following remark: if sampling weights are not properly defined, then the results given by any weighted procedures are questionable. This is one of the arguments in favour of using the proposed unweighted methods of chapter 3.

2.7 Analytical goals

Longitudinal surveys have many goals, including both descriptive and analytical inferences. Surveys have been used for descriptive purposes for hundreds of years. For example, governments have needed accurate pictures of the population in order to formulate new policies that would meet changing social circumstances. Inference from survey data is descriptive when the goal is to estimate quantities from a finite population $U = \{1, 2, \ldots, N\}$; such as, the population total $t_U = \sum_{i \in U} y_i$, or the population average $\bar{y}_U = t_U/N$. This area is well developed, see Cochran (1977) or Särndal et al. (1992) for further explanations and references.

In contrast, analytic use of survey data is more recent and has started to be used more frequently around the turn of the 20th century. One of the earliest examples of analytical inference from survey data is from the work of Snow (1855), who investigated the link between water supply and the incidence of cholera in the population of London in the mid-nineteenth century. Inference from survey data is analytical when the purpose is to make statements about an explanatory or causal mechanism of a probabilistic model, or about the parameters of a conceptual population or super-population. For example, when Statistics Canada carries out its monthly Labour Force Survey (LFS), the aim was to estimate the proportion (or number) of unemployed people in Canada; thus, the LFS serves descriptive purposes. In contrast, if the aim is to estimate the probability that a randomly selected individual in Canada will lose his/her job in some future month under similar conditions, then the purpose would be analytical (factors like age, sex and level of education can also be taken into account). An interesting alternative definition of analytical inference is given by Deming (1950), page 249. He describes analytic use of survey data as “directed at the underlying causes that have made the frequencies of various classes of the population what they are, and will govern the frequencies of these classes in time to come”. References for this area of sampling theory include Skinner et al. (1989) and Thompson (1997), chapter 6.

Analytical inference is very important in sample-based event history analysis which generally involves a probabilistic model, and where it is hard to imagine counterparts to parameters in a finite population context. Moreover, the real interest in event history analysis is generally to understand the different events that individuals experience over time and the factor influencing these events; not to estimate the parameters of a finite population. For example, in the Cox proportional hazards model given by

$$\lambda_i(t \mid x_i(t)) = \lambda_0(t)\ \exp\{\beta'\, x_i(t)\}\ , \tag{2.1}$$

the interest is in estimating β regardless of how the data is collected (i.e., through a longitudinal survey, with stratification and clustering, or through an i.i.d. sample). If β_U is defined as the solution of $U_W(\beta) = 0$, where $U_W(\beta)$ is given by

$$U_W(\beta) = \sum_{i \in U} \delta_i \left(x_i(t_i) - \frac{\sum_{j|t_j \geq t_i} x_j(t_i) \exp\{\beta'\, x_j(t_i)\}}{\sum_{j|t_j \geq t_i} \exp\{\beta'\, x_j(t_i)\}} \right)\ , \tag{2.2}$$

then its meaning as a descriptive parameter of the finite population U is clear. However, its more general meaning is not so clear, except for stochastic models which account for the heterogeneity in the population. Although β_U is well defined, there is no guarantee that it is close to β defined in (2.1). If they are not close, β_U and its estimator have little interest for analytical purposes. How close together β_U and β are depends on the stochastic model used, N and the sampling weights. This is similar to the linear regression problem discussed by Thompson (1997), pages 201–202, and Skinner et al. (1989), chapter 7.

For all these reasons, this thesis and, in particular, the methods proposed in chapter 3 are concerned with analytical inference from longitudinal surveys.

Chapter 3

Semi-parametric models and survey data

Statistical inference from data collected through longitudinal surveys, which involve the use of complex survey designs, must account for intra-cluster dependence and, sometimes, for response-selective sampling. If the goal is analytical inference for duration or survival time variables and if the sampling design is uninformative (see section 3.3), two classes of models that account for intra-cluster correlation are the proportional hazards frailty models and the marginal proportional hazards models. The methods proposed in this chapter are concerned with the latter. The focus is on semi-parametric proportional hazards models as introduced by Cox (1972). The specific goal is to extend the methods of Lee et al. (1992), Spiekerman & Lin (1998) and Lin et al. (2000) to the context of analytical inference carried out from longitudinal studies involving complex survey designs.

In this chapter, variance estimators for $\hat{\beta}$ and for the Breslow-Aalen estimator that account for intra-cluster dependence are proposed. Moreover, it is proven that, under mild conditions (see section 3.2.1), $\hat{\beta}$ is a consistent estimator and converges weakly to a zero-mean Gaussian process. The asymptotic properties of the Breslow-Aalen estimator for the baseline cumulative hazard function are also studied in the presence of intra-cluster correlation and it is shown that it converges weakly to a zero-mean Gaussian process, under the same mild conditions.

Section 3.1 presents the sampling design and statistical model on which the proposed methods are based. Section 3.2 describes the proposed methods in question, as summarized in the previous paragraph. Related theorems and proofs are also given. In section 3.3, uninformative and ignorable sampling designs, key assumptions made in section 3.2, are defined. In addition, the vital role these notions play in the application of the proposed methods to the context of longitudinal surveys is explained. Section 3.4 concludes with remarks on computer issues, on weighted versions of the methods proposed in section 3.2, and on similar methods such as Binder (1992), Lee et al. (1992), Spiekerman & Lin (1998), Lin (2000) and Lin et al. (2000).

3.1 Sampling design and model

Consider the following sampling design. First, the population U is divided into H disjoint strata U_h, $h = 1, \ldots, H$; that is, $U = \bigcup_{h=1}^{H} U_h$. Second, primary sampling units (PSU's) are selected according to a given sampling design (e.g., simple random sampling) within each stratum. Finally, within each PSU or cluster, observations are randomly selected according to a given sub-sampling design (e.g., systematic sampling). Thus, the sample S (of size n) can be expressed as

$$S = \bigcup_{h=1}^{H} \bigcup_{c=1}^{C_h} S_{hc} \, , \tag{3.1}$$

where S_{hc} is the sub-sample from the c^{th} cluster (or PSU) of the h^{th} stratum. Note that observations within a given cluster may be correlated, but observations from different clusters are not.

Let T_{hi} be the duration or survival time corresponding to the i^{th} individual of the h^{th} stratum. The marginal hazard functions of the T_{hi}'s are assumed to follow the stratified

Cox model

$$\lambda_{hi}(t) = \lambda_{0h}(t) \exp\{\beta_0' x_{hi}(t)\} \ , \tag{3.2}$$

where $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})'$ is a vector of unknown parameters and $x_{hi}(t)$ is a $p \times 1$ vector of external covariates for $i = 1, \ldots, n_h$ and $h = 1, \ldots, H$. As only the marginal distributions have been specified by (3.2), no assumption is made about the joint distribution for individuals in the same cluster.

3.2 Theoretical results

The approach taken in here is in the spirit of Lin & Wei (1989), Lee et al. (1992), Spiekerman & Lin (1998) and, in particular, Lin et al. (2000). That is, point estimates for β, for the cumulative hazard and for the survivor functions are obtained under the working assumption that the T_{hi}'s are mutually independent. However, the "robust" variance estimates remain valid when there is intra-cluster correlation. Note that, when there is intra-cluster correlation, martingale theory can not be used to derive asymptotic results for the marginal model given by (3.2). Instead, results from empirical process theory are employed; see Pollard (1990) and van der Vaart & Wellner (1996).

All theorems and proofs given in section 3.2 are derived under the sampling design and statistical model described in section 3.1, as well as under a few mild conditions given in section 3.2.1. In section 3.2.2 it is shown that $\hat{\beta}$ converges in probability to β_0. It is also proven, using estimating equation theory, that $C_{\bullet}^{1/2}(\hat{\beta} - \beta_0)$ is asymptotically normal with mean zero and covariance matrix that can be consistently estimated by $\widehat{V}_R(\hat{\beta}) = \mathcal{I}(\hat{\beta})^{-1} \left(\sum_{h=1}^{H} \sum_{c=1}^{C_h} \hat{U}_{hc}(\hat{\beta})\, \hat{U}_{hc}(\hat{\beta})' \right) \mathcal{I}(\hat{\beta})^{-1}$, a so-called "sandwich" variance estimator. The asymptotic properties of the Breslow-Aalen estimator for $\Lambda_{0h}(t)$ are discussed in section 3.2.3. It is shown that $C_h^{1/2}\big(\hat{\Lambda}_{0h}(t, \hat{\beta}) - \Lambda_{0h}(t)\big)$ converges weakly to a zero-mean Gaussian process, for $h = 1, \ldots, H$. Finally, variance estimators for the integrated hazard and survival functions that account for intra-cluster correlation are proposed in section 3.2.4.

3.2.1 Notation and assumptions

For $h = 1, \ldots, H$ and $i = 1, \ldots, n_h$, let t_{hi} denote the survival or censoring time corresponding to the i^{th} individual of the h^{th} stratum. Let δ_{hi} be the indicator that the observed value t_{hi} corresponds to a failure ($\delta_{hi} = 1$) or a censoring ($\delta_{hi} = 0$) time and let $x_{hi}(t)$ be the vector of time-varying covariates. The marginal distribution of T_{hi} is related to $x_{hi}(t)$ through model (3.2). Consider the sampling design described in section 3.1. It is convenient to introduce the following notation:

$$\begin{aligned}
\gamma_{hi}(t) &= I(t_{hi} \geq t) && (3.3)\\
S_h^{(0)}(\beta, t) &= \frac{1}{n_h} \sum_{i=1}^{n_h} \gamma_{hi}(t) \, \exp\{\beta' x_{hi}(t)\} && (3.4)\\
S_h^{(1)}(\beta, t) &= \frac{1}{n_h} \sum_{i=1}^{n_h} \gamma_{hi}(t) \, x_{hi}(t) \, \exp\{\beta' x_{hi}(t)\} \; , && (3.5)
\end{aligned}$$

for $i = 1, \ldots, n_h$ and $h = 1, \ldots, H$.

The following assumptions or conditions are assumed throughout chapter 3, for some constant $\tau > 0$:

1. The censoring mechanism is independent and non-informative, see section 1.2.4;

2. The sampling design $p(S|Z)$ is uninformative (see section 3.3) and each stratum sample S_h ($h = 1, \ldots, H$) can be partitioned into C_h disjoint groups or clusters S_{hc} of size n_{hc} ($c = 1, \ldots, C_h$). These clusters are such that observations from different clusters are independent, but observations within a cluster may not be. Note that this assumption allows the number of observations per cluster to vary from one cluster to another. This is less restrictive than Spiekerman & Lin (1998) (for example), where n_{hc} is assumed to be the same over all clusters;

3. For $i = 1, \ldots, n_h$ and $h = 1, \ldots, H$, $\Pr\{\gamma_{hi}(t) = 1\} > 0 \, \forall t \in [0, \tau)$; that is, attention is restricted to values of $t \in [0, \tau)$ such that the probability that any individual is at risk (i.e., alive and uncensored) is strictly positive for all t's;

4. $C_h^{1/2} \int_0^\tau (1 - J_h(t))\, \lambda_{0h}(t)\, dt \xrightarrow{\mathcal{P}} 0$, where $J_h(t) = I(\sum_{i=1}^{n_h} \gamma_{hi}(t) > 0)$;

5. There exist functions $s_h^{(r)}(\beta, t)$, $r = \{0, 1\}$, such that
$$\sup_{t\in[0,\tau),\beta\in\mathfrak{B}} \|S_h^{(r)}(\beta,t) - s_h^{(r)}(\beta,t)\| \xrightarrow{\mathcal{P}} 0 \quad \text{as } n_h \to \infty ,$$
for $h = 1, \ldots, H$ and where $\mathfrak{B}$ is a neighborhood of β_0;

6. $s_h^{(r)}(\beta, t)$ is a continuous function of β uniformly in t and is bounded on $\mathfrak{B} \times [0, \tau)$ for $r = \{0, 1\}$ and $h = 1, \ldots, H$;

7. $s_h^{(0)}(\beta, t)$ is bounded away from 0 on $\mathfrak{B} \times [0, \tau)$ for $h = 1, \ldots, H$, which implies that $S_h^{(0)}(\beta, t)$ is also bounded away from 0 in probability;

8. $\int_0^t \lambda_{0h}(u)\, du < \infty$ for all $t \in [0, \tau)$ and $h = 1, \ldots, H$;

9.
$$\Sigma_\beta = \sum_{h=1}^{H} \int_0^\tau \left(\frac{s_h^{(2)}(\beta_0,t)}{s_h^{(0)}(\beta_0,t)} - \left(\frac{s_h^{(1)}(\beta_0,t)}{s_h^{(0)}(\beta_0,t)}\right)\left(\frac{s_h^{(1)}(\beta_0,t)}{s_h^{(0)}(\beta_0,t)}\right)' \right) s_h^{(0)}(\beta_0,t)\, \lambda_{0h}(t)\, dt$$
is a symmetric $p \times p$ positive definite matrix, where $s_h^{(2)}(\beta,t) = \dfrac{\partial^2 s_h^{(0)}(\beta,t)}{\partial\beta^2}$.

Note that assumptions 6 and 7 imply that $s_h^{(1)}(\beta,t)/s_h^{(0)}(\beta,t)$ is bounded on $\mathfrak{B} \times [0,\tau)$ for $h = 1, \ldots, H$. In addition, these previous assumptions are fairly standard when studying the large sample properties of $\hat\beta$, $\hat\Lambda_{0h}(t, \hat\beta)$ and other related quantities under the stratified Cox model; see Andersen et al. (1993), conditions VII.2.1.

In summary, our asymptotic framework requires that the number of clusters per stratum $C_h \to \infty$ and that the number of observations per stratum $n_h \to \infty$. However, the number of strata H is considered fixed. Note that the case where all the observations in a given stratum are independent is covered by letting $C_h = n_h$ (i.e., $n_{hc} = 1$ for $c = 1, \ldots, C_h$).

3.2.2 Asymptotic properties of $\hat{\boldsymbol{\beta}}$

Estimation of β_0 in (3.2) is based on the estimating function given by (3.7) below. This function arises from the stratified Cox partial-likelihood with independent T_{hi}'s (e.g., Kalbfleisch & Prentice (2002), section 4.4, and Lawless (2003), section 7.1.6). It can be shown that (3.7) is an asymptotically unbiased estimating function, and hence valid for consistent estimation of β_0, when there is intra-cluster correlation.

Using the notation introduced in section 3.2.1, the partial-likelihood function corresponding to (3.2), under the independence working assumption, is

$$L(\beta) = \prod_{h=1}^{H} \prod_{i=1}^{n_h} \left(\frac{\exp\{\beta' x_{hi}(t_{hi})\}}{\sum_{j=1}^{n_h} \gamma_{hj}(t_{hi}) \exp\{\beta' x_{hj}(t_{hi})\}} \right)^{\delta_{hi}} , \tag{3.6}$$

where $\delta_{hi} = I(t_{hi}$ is a failure time). Taking the derivative with respect to β of the log of (3.6) yields the estimating function

$$U(\beta) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \delta_{hi} \left(x_{hi}(t_{hi}) - \frac{S_h^{(1)}(\beta, t_{hi})}{S_h^{(0)}(\beta, t_{hi})} \right) . \tag{3.7}$$

The estimator $\hat{\beta}$ is defined as the solution of $U(\beta) = 0$.

Using counting process notation (see section 1.3), (3.7) can be re-expressed as

$$n^{-1/2} U(\beta) = n^{-1/2} \sum_{h=1}^{H} \sum_{c=1}^{C_h} \sum_{i \in S_{hc}} \int_0^{\tau} \left(x_{hi}(t) - \frac{S_h^{(1)}(\beta, t)}{S_h^{(0)}(\beta, t)} \right) dN_{hi}(t) , \tag{3.8}$$

where $N_{hi}(t)$ was defined by (1.14) and $n = \sum_{h=1}^{H} n_h$ is the total number of observations. Let $d\Lambda_{hi}(t) = \gamma_{hi}(t) \exp\{\beta' x_{hi}(t)\}\, \lambda_{0h}(t)\, dt$ and note that $\Lambda_{hi}(t)$ is the compensator of $N_{hi}(t)$. Then, $M_{hi}(t) = N_{hi}(t) - \Lambda_{hi}(t)$ is a zero-mean local martingale on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ with respect to a filtration $\{\mathfrak{F}_{hit}; t \geq 0\}$, provided that $\gamma_{hi}(t)$ is predictable with respect to $\mathfrak{F}_{hit}$; see section 1.3.3. Then, (3.8) is equal to

$$n^{-1/2} U(\beta) = n^{-1/2} \sum_{h=1}^{H} \sum_{c=1}^{C_h} \sum_{i \in S_{hc}} \int_0^{\tau} \left(x_{hi}(t) - \frac{S_h^{(1)}(\beta, t)}{S_h^{(0)}(\beta, t)} \right) dM_{hi}(t) . \tag{3.9}$$

It is important to note that complications and additional technical challenges arise because, within clusters, observations are correlated. Even though the $M_{hi}(t)$'s are martingale with respect to the filtration $\{\mathcal{F}_{hit}; t \geq 0\}$, the term in brackets in (3.9) is not predictable with respect to this filtration. This is due to the fact that $S_h^{(r)}(\beta, t)$ (for $r = 0, 1$) depends on other individuals in the same cluster as the i^{th} individual. For the same reason, $\Lambda_{hi}(t)$ is not the compensator for $N_{hi}(t)$ given the history processes of other individuals included in the same cluster. This implies that $M_{hc\bullet}(t) = \sum_{i \in S_{hc}} M_{hi}(t)$ is not a martingale, but rather a zero-mean process (see (3.10) below). In addition, this also implies that (3.9) is only asymptotically unbiased in the presence of intra-cluster correlation (see note following (3.18)). Nevertheless, to use the theory of estimating equation to derive a robust or "sandwich" variance estimator for $\hat{\beta}$, it is necessary to express (3.9) as asymptotically the sum of independent random vectors. Fortunately, this can still be done using properties of zero-mean processes combined with empirical process theory. To this end, first note that (3.9) can be re-written as

$$\begin{aligned} n^{-1/2}U(\beta) = n^{-1/2} &\sum_{h=1}^{H} \sum_{c=1}^{C_h} \sum_{i \in S_{hc}} \int_0^{\tau} \left(x_{hi}(t) - \frac{s_h^{(1)}(\beta, t)}{s_h^{(0)}(\beta, t)} \right) dM_{hi}(t) \\ &- n^{-1/2} \sum_{h=1}^{H} \sum_{c=1}^{C_h} \int_0^{\tau} \left(\frac{S_h^{(1)}(\beta, t)}{S_h^{(0)}(\beta, t)} - \frac{s_h^{(1)}(\beta, t)}{s_h^{(0)}(\beta, t)} \right) dM_{hc\bullet}(t) \ . \end{aligned} \tag{3.10}$$

Second, the term inside the second integral in (3.10) can be expressed as

$$\begin{aligned} \frac{S_h^{(1)}(\beta, t)}{S_h^{(0)}(\beta, t)} - \frac{s_h^{(1)}(\beta, t)}{s_h^{(0)}(\beta, t)} = &\frac{S_h^{(1)}(\beta, t) - s_h^{(1)}(\beta, t)}{s_h^{(0)}(\beta, t)} \\ &+ \left(\frac{1}{S_h^{(0)}(\beta, t)} - \frac{1}{s_h^{(0)}(\beta, t)} \right) \left(S_h^{(1)}(\beta, t) - s_h^{(1)}(\beta, t) \right) \\ &- \left(S_h^{(0)}(\beta, t) - s_h^{(0)}(\beta, t) \right) \frac{s_h^{(1)}(\beta, t)}{\left(s_h^{(0)}(\beta, t) \right)^2} \\ &- \left(\frac{1}{S_h^{(0)}(\beta, t)} - \frac{1}{s_h^{(0)}(\beta, t)} \right) \left(S_h^{(0)}(\beta, t) - s_h^{(0)}(\beta, t) \right) \frac{s_h^{(1)}(\beta, t)}{s_h^{(0)}(\beta, t)} \ , \end{aligned} \tag{3.11}$$

for $h = 1, \ldots, H$.

Third, by the assumptions of section 3.2.1,

$$S_h^{(r)}(\beta,t) - s_h^{(r)}(\beta,t) = \mathrm{o}_P(1) \tag{3.12}$$

and

$$\frac{1}{S_h^{(0)}(\beta,t)} - \frac{1}{s_h^{(0)}(\beta,t)} = \mathrm{o}_P(1) \ , \tag{3.13}$$

uniformly in t for $r = \{0,1\}$ and $h = 1,\ldots,H$.

Fourth,

$$C_h^{-1/2} \sum_{c=1}^{C_h} M_{hc\bullet}(t) \tag{3.14}$$

is the sum of independent zero-mean terms for fixed t. Under the assumptions of section 3.2.1 and a few mild assumptions on the variances of the $M_{hc\bullet}(t)$'s, it is readily shown that (3.14) converges in finite dimensional distribution to a zero-mean Gaussian process by the central limit theorem as $C_h \to \infty$ for each $h = 1,\ldots,H$. In addition, $M_{hc\bullet}(t)$ is the difference of two monotone functions in t and, thus, is manageable; see Pollard (1990), page 38. Therefore, by the functional central limit theorem, (3.14) converges weakly to a zero-mean Gaussian process. References on the functional central limit theorem include Pollard (1990), page 53, and van der Vaart & Wellner (1996), section 2.11.

Let $C_\bullet = \sum_{h=1}^{H} C_h$, and note that

$$C_\bullet^{-1/2} \sum_{h=1}^{H} \sum_{c=1}^{C_h} M_{hc\bullet}(t) \tag{3.15}$$

is a linear combination of zero-mean Gaussian processes. This combined with the arguments and reasoning given in the previous paragraph, implies that (3.15) converges weakly to a zero-mean Gaussian process as $\min C_h \to \infty$. Moreover, by Serfling (1980), page 8,

$$C_\bullet^{-1/2} \sum_{h=1}^{H} \sum_{c=1}^{C_h} M_{hc\bullet}(t) = \mathrm{O}_P(1) \ . \tag{3.16}$$

Other references on O_P, o_P and stochastic sequences include Bishop et al. (1975), section 14.4.

Finally, combining (3.11)–(3.16), the second term in (3.10) is of order $o_P(\sqrt{C_\bullet/n})$. Therefore, the partial-score function given by (3.9) can be re-expressed as

$$n^{-1/2}U(\beta) = n^{-1/2}\sum_{h=1}^{H}\sum_{c=1}^{C_h} U_{hc}(\beta) + o_p\left(\sqrt{C_\bullet/n}\right) \ , \tag{3.17}$$

where the $U_{hc}(\beta)$'s are independent terms given by

$$U_{hc}(\beta) = \sum_{i\in S_{hc}} \int_0^\tau \left(x_{hi}(t) - \frac{s_h^{(1)}(\beta,t)}{s_h^{(0)}(\beta,t)} \right) dM_{hi}(t) \ . \tag{3.18}$$

Note that the terms in brackets in (3.18) are deterministic and that $E(dM_{hi}(t)) = 0$ for each h, i combination; thus, $E(U_{hc}(\beta)) = 0$. Combining this with (3.17) shows that $U(\beta)$, given by (3.7), is asymptotically unbiased.

From (3.17) and estimating equations theory, a robust or "sandwich" variance estimator for $\hat\beta$ is

$$\widehat{V}_R(\hat\beta) = \mathcal{I}(\hat\beta)^{-1}\left(\sum_{h=1}^{H}\sum_{c=1}^{C_h} \hat{U}_{hc}(\hat\beta)\hat{U}_{hc}(\hat\beta)'\right)\mathcal{I}(\hat\beta)^{-1} \ , \tag{3.19}$$

where $\mathcal{I}(\beta) = -\partial U(\beta)/\partial\beta'$ and

$$\begin{aligned}\hat{U}_{hc}(\hat\beta) = &\sum_{i\in S_{hc}} \delta_{hi}\left(x_{hi}(t_{hi}) - \frac{S_h^{(1)}(\hat\beta, t_{hi})}{S_h^{(0)}(\hat\beta, t_{hi})} \right) \\ &- \sum_{i\in S_{hc}}\sum_{j=1}^{n_h} \left(x_{hi}(t_{hj}) - \frac{S_h^{(1)}(\hat\beta, t_{hj})}{S_h^{(0)}(\hat\beta, t_{hj})} \right) \frac{\delta_{hj}\gamma_{hi}(t_{hj})\exp\{\hat\beta' x_{hi}(t_{hj})\}}{n_h S_h^{(0)}(\hat\beta, t_{hj})} \ .\end{aligned} \tag{3.20}$$

The properties of $\hat\beta$ and $\widehat{V}_R(\hat\beta)$ are given in theorems 3.1–3.3.

Theorem 3.1 (Consistency of $\hat{\boldsymbol\beta}$) Under the assumptions of section 3.2.1, $\hat\beta$ is a consistent estimator of β_0. □

Proof

The consistency of $\hat\beta$ can be proven following the same ideas and steps as in Lin et al. (2000), appendix A.1. Since section 3.2 is based on the stratified Cox model, given by (3.2), small modifications must be made to their proof. To this end, consider the log-likelihood function $l(\beta)$, where $l(\beta) = \log L(\beta)$ and $L(\beta)$ was defined in (3.6). Using counting process notation, it can be expressed as

$$l(\beta) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \left(\int_0^\tau \beta' x_{hi}(t) - \log\left(n_h S_h^{(0)}(\beta, t) \right) \right) dN_{hi}(t) \; . \tag{3.21}$$

Define $X(\beta) = n^{-1}\Big(l(\beta) - l(\beta_0) \Big)$; that is,

$$X(\beta) = n^{-1} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \int_0^\tau \left[(\beta - \beta_0)' x_{hi}(t) dN_{hi}(t) - \log\left(\frac{S_h^{(0)}(\beta, t)}{S_h^{(0)}(\beta_0, t)} \right) \right] dN_{hi}(t) \; . \tag{3.22}$$

It can be shown that $X(\beta) \to \mathcal{X}(\beta)$ almost surely for every β; see Lin et al. (2000), appendix A.1. In addition, it is readily shown that $\partial^2 \mathcal{X}(\beta)/\partial \beta^2|_{\beta=\beta_0} = -\Sigma_\beta$, where Σ_β is positive definite by assumption 9 of section 3.2.1. This and the fact that $\partial \mathcal{X}(\beta)/\partial\beta|_{\beta=\beta_0} = 0$ prove that $\mathcal{X}(\beta)$ is concave. To show that $X(\beta)$ is also concave, compute its second derivative with respect to β; that is,

$$\begin{aligned} \frac{\partial^2 X(\beta)}{\partial \beta^2} = n^{-1} \sum_{h=1}^{H} \Bigg[\int_0^\tau & \frac{\sum_{i=1}^{n_h} \gamma_{hi}(t) x_{hi}(t)^{\otimes 2} \exp\{\beta' \, x_{hi}(t)\}}{n_h \, S_h^{(0)}(\beta, t)} \\ & - \left(\frac{S_h^{(1)}(\beta, t)}{S_h^{(0)}(\beta, t)} \right)^{\otimes 2} dN_{h\bullet}(t) \Bigg] \; , \end{aligned} \tag{3.23}$$

where $N_{h\bullet}(t) = \sum_{i=1}^{n_h} N_{hi}(t)$ and $v^{\otimes 2} = vv'$. With a few simple algebraic manipulations (e.g., multiplying the second term in the above integral by $S_h^{(0)}/S_h^{(0)}$), (3.23) can be re-expressed as

$$\frac{\partial^2 X(\beta)}{\partial \beta^2} = \frac{-1}{n \, n_h} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \int_0^\tau \left(x_{hi}(t) - \frac{S_h^{(1)}(\beta, t)}{S_h^{(0)}(\beta, t)} \right)^{\otimes 2} \frac{\gamma_{hi}(t) exp\{\beta' \, x_{hi}(t)\}}{S_h^{(0)}(\beta, t)} \, dN_{h\bullet}(t) \; , \tag{3.24}$$

which is negative semidefinite, and thus $X(\beta)$ is concave. This implies that the convergence of $X(\beta)$ to $\mathcal{X}(\beta)$ is uniform on any compact set of β; see Rockafellar (1970), theorem 1.8. The few remaining arguments required to show that $l(\beta)$ has a unique maximizer $\hat{\beta}$, given as the solution to the estimating equation $U(\beta) = 0$, are practically identical to the ones given by Lin et al. (2000) in the last paragraph of their appendix A.1; they are thus omitted. QED

The following lemma is needed in the proof theorem 3.2.

Lemma 3.1 (Asymptotic normality of $\boldsymbol{U(\beta)}$) Under the assumptions of section 3.2.1 and mild assumptions on the variances of the $U_{hc}(\beta)$'s,

$$C_\bullet^{-1/2}\, U(\beta) \xrightarrow{\mathcal{D}} N_p\left(0, \Sigma_U\right) \ ,$$

where

$$\Sigma_U = \operatorname{Var}\left(\frac{1}{C_\bullet^{1/2}} U(\beta)\right) = \frac{1}{C_\bullet} \sum_{h=1}^{H} \sum_{c=1}^{C_h} \sum_{c'=1}^{C_h} E\left(\widetilde{U}_{hc}(\beta)\, \widetilde{U}_{hc'}(\beta)'\right) \tag{3.25}$$

and

$$\widetilde{U}_{hc}(\beta) = \sum_{i \in S_{hc}} \int_0^\tau \left(x_{hi}(t) - \frac{S_h^{(1)}(\beta,t)}{S_h^{(0)}(\beta,t)} \right) dM_{hi}(t) \ . \tag{3.26}$$

□

Contrary to the $U_{hc}(\beta)$'s given by (3.18), the $\widetilde{U}_{hc}(\beta)$'s are not independent random vectors.

Proof

Looking at (3.17), $U(\beta)$ is asymptotically a sum of $C_\bullet$ independent random vectors with mean 0; recall that $E(U_{hc}(\beta)) = 0$, where $U_{hc}(\beta)$ is given by (3.18). Under mild conditions on the variances of these $U_{hc}(\beta)$'s, the multivariate central limit theorem implies weak convergence to a p-variate normal distribution. QED

Theorem 3.2 (Asymptotic normality of $\hat{\boldsymbol{\beta}}$) Under the assumptions of section 3.2.1,

$$C_\bullet^{1/2}\, (\hat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} N_p\big(0, \Sigma\big) \ ,$$

where $\Sigma = \Sigma_\beta^{-1} \Sigma_U \Sigma_\beta^{-1}$ and Σ_β was given in assumption 9 of section 3.2.1. □

Proof

A Taylor series expansion of $U(\beta)$ around β_0 yields,

$$U(\beta) = U(\beta_0) - \mathcal{I}(\beta^\star)\, (\beta - \beta_0) \ ,$$

where $\mathcal{I}(\beta) = -\, \partial U(\beta)/\partial \beta'$ and $\beta^\star$ is on the line segment between β_0 and $\hat{\beta}$. Since $U(\hat{\beta}) = 0$, replacing β by $\hat{\beta}$ yields

$$C_\bullet^{1/2}\, (\hat{\beta} - \beta_0) = \left(\frac{\mathcal{I}(\beta^\star)}{C_\bullet}\right)^{-1} C_\bullet^{-1/2}\, U(\beta_0) \ . \tag{3.27}$$

The convergence of $C_\bullet^{-1}\mathcal{I}(\beta^\star)$ to Σ_β is easily proven using the consistency of $\hat\beta$ and theorem 2 of Spiekerman & Lin (1998). Combining the convergence of $C_\bullet^{-1}\mathcal{I}(\beta^\star)$, lemma 3.1 and (3.27) yield theorem 3.2. QED

Natural estimators of Σ_β and Σ_U are, respectively,

$$\hat\Sigma_\beta = \frac{1}{C_\bullet}\,\mathcal{I}(\hat\beta)$$

and

$$\hat\Sigma_U = \frac{1}{C_\bullet}\sum_{h=1}^{H}\sum_{c=1}^{C_h}\hat U_{hc}(\hat\beta)\,\hat U_{hc}(\hat\beta)'\ ,$$

where $\hat U_{hc}(\hat\beta)$ was given in (3.20).

Lemma 3.2 (Consistency of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{U}}$) Under the assumptions of section 3.2.1, $\hat\Sigma_U$ is a consistent estimator of Σ_U. □

Proof

First, rewrite $\hat U_{hc}(\hat\beta)$, given by (3.20), using counting process notation; that is,

$$\hat U_{hc}(\hat\beta) = \sum_{i\in S_{hc}}\int_0^\tau\left(x_{hi}(t) - \frac{S_h^{(1)}(\hat\beta,t)}{S_h^{(0)}(\hat\beta,t)}\right)dM_{hi}(t)\ . \tag{3.28}$$

Combining that $\hat\beta \xrightarrow{\mathcal{P}} \beta_0$ (see theorem 3.1) with (3.12) and (3.13) it is readily shown that

$$\frac{1}{C_\bullet}\sum_{h=1}^{H}\sum_{c=1}^{C_h}\hat U_{hc}(\hat\beta)\,\hat U_{hc}(\hat\beta)' - \frac{1}{C_\bullet}\sum_{h=1}^{H}\sum_{c=1}^{C_h}U_{hc}(\beta)\,U_{hc}(\beta)' \xrightarrow{\mathcal{P}} 0\ . \tag{3.29}$$

A formal proof of (3.29) is obtained by applying the same ideas as the ones used to go from (3.10) to (3.17) to each element of the symmetric $p\times p$ matrix given in (3.29).

From the weak law of large numbers

$$\frac{1}{C_\bullet}\sum_{h=1}^{H}\sum_{c=1}^{C_h}U_{hc}(\beta)\,U_{hc}(\beta)' - \frac{1}{C_\bullet}\sum_{h=1}^{H}\sum_{c=1}^{C_h}E\left(U_{hc}(\beta)\,U_{hc}(\beta)'\right) \xrightarrow{\mathcal{P}} 0\ ; \tag{3.30}$$

see Serfling (1980), section 1.8. Recall that the $U_{hc}(\beta)$'s, given by (3.18), are independent random vectors.

From (3.17) and the properties of stochastic orders,

$$C_\bullet^{-1/2}\, U(\beta) = C_\bullet^{-1/2} \sum_{h=1}^{H} \sum_{c=1}^{C_h} \widetilde{U}_{hc}(\beta) = C_\bullet^{-1/2} \sum_{h=1}^{H} \sum_{c=1}^{C_h} U_{hc}(\beta) + o_P(1) \;, \tag{3.31}$$

where $C_\bullet^{-1/2} U(\beta) \xrightarrow{\mathcal{D}} N_p\left(0, \Sigma_U\right)$; see lemma 3.1. However, (3.31) combined with Bishop et al. (1975), theorem 14.4-3 (that is, if $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n = X_n + o_p(1)$ then $Y_n \xrightarrow{\mathcal{D}} X$), implies that

$$C_\bullet^{-1/2} \sum_{h=1}^{H} \sum_{c=1}^{C_h} U_{hc}(\beta) \xrightarrow{\mathcal{D}} N_p\left(0, \Sigma_U\right) \;. \tag{3.32}$$

Therefore, an alternative definition of Σ_U is given by

$$\mathrm{Var}\Bigg(C_\bullet^{-1/2} \sum_{h=1}^{H} \sum_{c=1}^{C_h} U_{hc}(\beta)\Bigg) = \frac{1}{C_\bullet} \sum_{h=1}^{H} \sum_{c=1}^{C_h} E\left(U_{hc}(\beta)\, U_{hc}(\beta)'\right) \tag{3.33}$$

Combining (3.29), (3.30) and (3.33) yield the consistency of $\hat{\Sigma}_U$. QED

Theorem 3.3 (Consistency of $\widehat{V}_{R}(\hat{\beta})$ and $\hat{\Sigma}$) Under the assumptions of section 3.2.1,

$$\hat{\Sigma} = C_\bullet \widehat{V}_R(\hat{\beta}) = \hat{\Sigma}_\beta^{-1} \hat{\Sigma}_U \hat{\Sigma}_\beta^{-1}$$

is a consistent estimator of Σ. □

Proof

The proof of theorem 3.3 is based on the convergence of $\hat{\Sigma}_\beta \xrightarrow{\mathcal{P}} \Sigma_\beta$ and of $\hat{\Sigma}_U \xrightarrow{\mathcal{P}} \Sigma_U$. The consistency of $\hat{\Sigma}_\beta$ follows from the consistency of $\hat{\beta}$ (see theorem 3.1) and from the consistency of $C_\bullet^{-1}\, \mathcal{I}(\beta^\star)$ (see theorem 3.2). Finally, the consistency of $\hat{\Sigma}_U$ is a consequence of lemma 3.2. QED

3.2.3 Asymptotic properties of the Breslow-Aalen estimator

Under the assumptions of section 3.2.1 and the stratified proportional hazards model given by (3.2), the Breslow-Aalen estimator is defined as

$$\hat{\Lambda}_{0h}(t, \hat{\beta}) = \int_0^t \frac{J_h(u)}{n_h S_h^{(0)}(\hat{\beta}, u)}\, dN_{h\bullet}(u) \quad \text{for } h = 1, \ldots, H \;, \tag{3.34}$$

where $N_{h\bullet}(t) = \sum_{i=1}^{n_h} N_{hi}(t)$ and $J_h(t) = I\left(\sum_{i=1}^{n_h} \gamma_{hi}(t) > 0\right)$, with the convention that $0/0 \equiv 0$. Defining

$$\Lambda_{0h}(t) = \int_0^t \lambda_{oh}(u)\, du \ ,$$

the asymptotic distribution of $\hat{\Lambda}_{0h}(t, \hat{\beta})$ is given by the following theorem.

Theorem 3.4 Under the assumptions of section 3.2.1, the process

$$C_h^{1/2}\Big(\hat{\Lambda}_{0h}(t,\hat{\beta}) - \Lambda_{0h}(t)\Big) \qquad 0 \le t < \tau$$

converges weakly to a Gaussian process with mean 0 and covariance function

$$\begin{aligned} \sigma_\Lambda(t,s) = C_h \sum_{c=1}^{C_h} \sum_{c'=1}^{C_h} E\Big(\psi_{hc}(t)\psi_{hc'}(s)\Big) \\ + C_h\, g_h(\beta_0,t)'\, \frac{\Sigma_\beta^{-1}}{C_\bullet} \left(\sum_{h' \neq h} \sum_{c=1}^{C_{h'}} \mathrm{Var}\Big(U_{h'c}(\beta_0)\Big) \right) \frac{\Sigma_\beta^{-1}}{C_\bullet}\, g_h(\beta_0,s) \ , \end{aligned} \tag{3.35}$$

where

$$\psi_{hc}(t) = \int_0^t \frac{1}{n_h s_h^{(0)}(\beta_0,u)}\, dM_{hc\bullet}(u) - g_h(\beta_0,t)'\, \frac{\Sigma_\beta^{-1}}{C_\bullet}\, U_{hc}(\beta_0) \ , \tag{3.36}$$

$$g_h(\beta,t) = \int_0^t \frac{s_h^{(1)}(\beta,u)}{s_h^{(0)}(\beta,u)}\, \lambda_{0h}(u)\, du \ , \tag{3.37}$$

and where $U_{hc}(\beta)$ and Σ_β are, respectively, defined by (3.18) and in assumption 9 of section 3.2.1. In addition, $\sigma_\Lambda(t,s)$ is consistently estimated, uniformly in t and s, by

$$\begin{aligned} \hat{\sigma}_\Lambda(t,s) = C_h \sum_{c=1}^{C_h} \hat{\psi}_{hc}(t)\hat{\psi}_{hc}(s) \\ + C_h\, G_h(\hat{\beta},t)'\, \mathcal{I}^{-1}(\hat{\beta}) \left(\sum_{h' \neq h} \sum_{c=1}^{C_{h'}} \hat{U}_{h'c}(\hat{\beta})\, \hat{U}_{h'c}(\hat{\beta})' \right) \mathcal{I}^{-1}(\hat{\beta})\, G_h(\hat{\beta},s) \ , \end{aligned} \tag{3.38}$$

where $\mathcal{I}(\beta) = -\partial U(\beta)/\partial \beta'$, $\hat{U}_{hc}(\hat{\beta})$ is defined by (3.20),

$$\hat{\psi}_{hc}(t) = \sum_{i \in S_{hc}} \int_0^t \frac{1}{n_h S_h^{(0)}(\hat{\beta},u)}\, d\hat{M}_{hi}(u) - G_h(\hat{\beta},t)'\, \mathcal{I}^{-1}(\hat{\beta})\, \hat{U}_{hc}(\hat{\beta}) \ , \tag{3.39}$$

$$\hat{M}_{hi}(t) = N_{hi}(t) - \int_0^t \gamma_{hi}(u)\, \exp\{\hat{\beta}' x_{hi}(u)\}\, d\hat{\Lambda}_{0h}(u,\hat{\beta}) \ , \tag{3.40}$$

and

$$G_h(\beta,t) = \int_0^t \frac{S_h^{(1)}(\beta,u)}{S_h^{(0)}(\beta,u)} \, \frac{J_h(u)}{n_h S_h^{(0)}(\beta,u)} \, dN_{h\bullet}(u) \ . \tag{3.41}$$

□

Note that the middle term of the second term of (3.38) shares strong similarities with $\widehat{V}_R(\hat{\beta})$, given by (3.19); the only difference being that in the former the summation does not include observations from the h^{th} stratum.

Proof

To prove the first part of theorem 3.4, first write $\hat{\Lambda}_{0h}(t,\hat{\beta}) - \Lambda_{0h}(t)$ as

$$\begin{aligned} C_h^{1/2}\Big(\hat{\Lambda}_{0h}(t,\hat{\beta}) - \Lambda_{0h}(t)\Big) &= C_h^{1/2}\Big(\hat{\Lambda}_{0h}(t,\hat{\beta}) - \hat{\Lambda}_{0h}(t,\beta_0)\Big) \\ &\quad + C_h^{1/2}\Big(\hat{\Lambda}_{0h}(t,\beta_0) - \Lambda_{0h}(t)\Big) \ . \end{aligned} \tag{3.42}$$

The two terms on the right hand side of (3.42) are dealt with separately. The second term can be re-expressed as

$$\begin{aligned} C_h^{1/2}\Big(\hat{\Lambda}_{0h}(t,\beta_0) - \Lambda_{0h}(t)\Big) &= C_h^{1/2} \int_0^t \frac{J_h(u)}{n_h S_h^{(0)}(\beta_0,u)} \, dM_{h\bullet}(u) \\ &\quad + C_h^{1/2} \int_0^t \big(J_h(u) - 1\big) \, \lambda_{0h}(u) \, du \ , \end{aligned} \tag{3.43}$$

where $M_{h\bullet}(t) = N_{h\bullet}(t) - \Lambda_{0h}(t)$. As mentioned previously, due to the correlation between the observations of the h^{th} stratum, $M_{h\bullet}(t)$ is not a martingale, but a zero-mean process.

By assumption 4 of section 3.2.1, the last term on the right hand side of (3.43) is of order $o_P(1)$, and (3.43) can be re-written as

$$\begin{aligned} C_h^{1/2}\Big(\hat{\Lambda}_{0h}(t,\beta_0) - \Lambda_{0h}(t)\Big) &= C_h^{1/2} \int_0^t J_h(u) \left(\frac{1}{S_h^{(0)}(\beta_0,u)} - \frac{1}{s_h^{(0)}(\beta_0,u)} \right) \frac{dM_{h\bullet}(u)}{n_h} \\ &\quad + C_h^{1/2} \int_0^t \frac{J_h(u)}{n_h s_h^{(0)}(\beta_0,u)} \, dM_{h\bullet}(u) + o_P(1) \ . \end{aligned} \tag{3.44}$$

To show that the first term on the right hand side of (3.44) is of oder $o_p(\sqrt{C_h/n_h})$, first note that

$$n_h^{-1/2} M_{h\bullet}(t) = n_h^{-1/2} \sum_{i=1}^{n_h} M_{hi}(t) = \left(\frac{C_h}{n_h}\right)^{1/2} C_h^{-1/2} \sum_{c=1}^{C_h} M_{hc\bullet}(t) \ , \tag{3.45}$$

where the second term on the right hand side of the last equality is (3.14), which was shown to converge weakly to a zero-mean Gaussian process on page 60. Hence,

$$n_h^{-1/2} M_{h\bullet}(t) \tag{3.46}$$

converges weakly to a zero-mean Gaussian process, under the assumption that $C_h/n_h \to$ constant as C_h and $n_h \to \infty$. Combining the weak convergence of (3.46) with (3.13) and the fact that $J_h(t) \xrightarrow{\mathcal{P}} 1$, uniformly in t, yields that

$$\frac{C_h^{1/2}}{n_h^{1/2}} \int_0^t J_h(u) \left(\frac{1}{S_h^{(0)}(\beta_0, u)} - \frac{1}{s_h^{(0)}(\beta_0, u)} \right) \frac{dM_{h\bullet}(u)}{n_h^{1/2}} = o_P\left(\sqrt{C_h/n_h}\right) \ , \tag{3.47}$$

uniformly in t. Finally, combining (3.44) and (3.47), the second term on the right hand side of (3.42) can be written as

$$C_h^{1/2}\Big(\hat{\Lambda}_{0h}(t,\beta_0) - \Lambda_{0h}(t)\Big) = C_h^{1/2} \int_0^t \frac{J_h(u)}{n_h s_h^{(0)}(\beta_0, u)} \, dM_{h\bullet}(u) + o_P(1) \tag{3.48}$$

since $C_h \leq n_h \Rightarrow \sqrt{C_h/n_h} \leq 1$.

Using Taylor series expansion around β_0, the first term on the right hand side of (3.42) can be re-expressed as

$$C_h^{1/2}\Big(\hat{\Lambda}_{0h}(t,\hat{\beta}) - \hat{\Lambda}_{0h}(t,\beta_0)\Big) = -C_h^{1/2} G_h(\beta^\star, t)' \, (\hat{\beta} - \beta_0) \ , \tag{3.49}$$

where $G_h(\beta, t)$ is given by (3.41), and $\beta^\star$ is on the line segment between $\hat{\beta}$ and β_0.

Next note that the difference between $G_h(\beta, t)$ and $g_h(\beta, t)$, defined respectively by (3.41) and (3.37), can be expressed as

$$\begin{aligned} G_h(\beta,t) - g_h(\beta,t) = & \int_0^t \left(\frac{S_h^{(1)}(\beta,u)}{(S_h^{(0)}(\beta,u))^2} - \frac{s_h^{(1)}(\beta,u)}{(s_h^{(0)}(\beta,u))^2} \right) J_h(u) \, \frac{dM_{h\bullet}(u)}{n_h} \\ & + \int_0^t \frac{s_h^{(1)}(\beta,u)}{(s_h^{(0)}(\beta,u))^2} \, J_h(u) \, \frac{dM_{h\bullet}(u)}{n_h} \\ & + \int_0^t J_h(u) \left(\frac{S_h^{(1)}(\beta,u)}{S_h^{(0)}(\beta,u)} - \frac{s_h^{(1)}(\beta,u)}{s_h^{(0)}(\beta,u)} \right) \lambda_{0h}(u) \, du \\ & + \int_0^t \frac{s_h^{(1)}(\beta,u)}{s_h^{(0)}(\beta,u)} \, (J_h(u) - 1) \, \lambda_{0h}(u) \, du \ . \end{aligned} \tag{3.50}$$

Assumptions 4, 6 and 7 of section 3.2.1 imply that the last term on the right hand side of (3.50) is of order $o_P(1)$. Similarly, (3.11)–(3.13) and the fact that $J_h(t) \xrightarrow{\mathcal{P}} 1$ imply that the third term on the right hand side of (3.50) is of order $o_P(1)$. The weak convergence of (3.46) and the properties of stochastic orders (see Bishop et al. (1975), page 478) yield that $n_h^{-1} M_{h\bullet}(t) = o_P(1)$. Combining this with the fact that, by assumptions, $s_h^{(1)}(\beta, t)/(s_h^{(0)}(\beta, t))^2$ is bounded on $[0, \tau)$ and that $J_h(t) \xrightarrow{\mathcal{P}} 1$, the second term on the right hand side of (3.50) is of order $o_P(1)$ as well. Finally, the inside of the first on the right hand side of (3.50) can be re-written as

$$\begin{aligned} \frac{S_h^{(1)}(\beta, u)}{(S_h^{(0)}(\beta, u))^2} - \frac{s_h^{(1)}(\beta, u)}{(s_h^{(0)}(\beta, u))^2} = {} & \frac{S_h^{(1)}(\beta, u) - s_h^{(1)}(\beta, u)}{(s_h^{(0)}(\beta, u))^2} \\ & + \left(\frac{1}{(S_h^{(0)}(\beta, u))^2} - \frac{1}{(s_h^{(0)}(\beta, u))^2} \right) s_h^{(1)}(\beta, u) \qquad (3.51) \\ & + \left(S_h^{(1)}(\beta, u) - s_h^{(1)}(\beta, u) \right) \left(\frac{1}{(S_h^{(0)}(\beta, u))^2} - \frac{1}{(s_h^{(0)}(\beta, u))^2} \right) . \end{aligned}$$

Hence, combining (3.51), (3.12), (3.13), the convergence of $J_h(t)$ and the fact that $n_h^{-1} M_{h\bullet}(t) = o_P(1)$ implies that the first term on the right hand side of (3.50) is, like the other 3 terms, of order $o_P(1)$.

Since $\hat{\beta} \xrightarrow{\mathcal{P}} \beta_0 \Rightarrow G_h(\beta^\star, t) \xrightarrow{\mathcal{P}} G_h(\beta_0, t)$ and (3.50) is of order $o_P(1)$, it is now possible to re-write (3.49) as

$$C_h^{1/2} \left(\hat{\Lambda}_{0h}(t, \hat{\beta}) - \hat{\Lambda}_{0h}(t, \beta_0) \right) = -C_h^{1/2} g_h(\beta_0, t)'(\hat{\beta} - \beta_0) - C_h^{1/2} (\hat{\beta} - \beta_0)\, o_P(1) \ . \qquad (3.52)$$

However, theorem 3.2 implies that $C_h^{1/2}(\hat{\beta} - \beta_0) = O_P(1)$. Hence, the last term on the right hand side of (3.52) is of order $o_P(1)$.

Combining (3.48) and (3.52), (3.42) becomes

$$\begin{aligned} C_h^{1/2} \left(\hat{\Lambda}_{0h}(t, \hat{\beta}) - \Lambda_{0h}(t) \right) = {} & C_h^{1/2} \int_0^t \frac{J_h(u)}{n_h s_h^{(0)}(\beta_0, u)} \, dM_{h\bullet}(u) \\ & - C_h^{1/2} g_h(\beta_0, t)'(\hat{\beta} - \beta_0) + o_P(1) \ . \end{aligned} \qquad (3.53)$$

Using (3.27), the fact that $C_\bullet^{-1} \mathcal{I}(\beta^\star) \xrightarrow{\mathcal{P}} \Sigma_\beta$ (see theorem 3.2) and (3.17), (3.53) can be re-expressed as

$$\begin{aligned} C_h^{1/2}\Big(\hat{\Lambda}_{0h}(t,\hat{\beta}) - \Lambda_{0h}(t)\Big) = {} & C_h^{1/2} \int_0^t \frac{J_h(u)}{n_h s_h^{(0)}(\beta_0,u)}\, dM_{h\bullet}(u) \\ & - C_h^{1/2} g_h(\beta_0,t)' \frac{\Sigma_\beta^{-1}}{C_\bullet} \left(\sum_{h=1}^{H} \sum_{c=1}^{C_h} U_{hc}(\beta_0) \right) + o_P(1) \ , \end{aligned} \tag{3.54}$$

where $U_{hc}(\beta)$ is given by (3.18). Putting together all the observations from stratum h, (3.54) becomes

$$\begin{aligned} C_h^{1/2}\Big(\hat{\Lambda}_{0h}(t,\hat{\beta}) - \Lambda_{0h}(t)\Big) = {} & C_h^{1/2} \sum_{c=1}^{C_h} \int_0^t \frac{J_h(u)}{n_h s_h^{(0)}(\beta_0,u)}\, dM_{hc\bullet}(u) \\ & - C_h^{1/2} g_h(\beta_0,t)' \frac{\Sigma_\beta^{-1}}{C_\bullet} \left(\sum_{c=1}^{C_h} U_{hc}(\beta_0) \right) \\ & - C_h^{1/2} g_h(\beta_0,t)' \frac{\Sigma_\beta^{-1}}{C_\bullet} \left(\sum_{h' \neq h} \sum_{c=1}^{C_h} U_{hc}(\beta_0) \right) + o_P(1) \ , \end{aligned} \tag{3.55}$$

where $M_{hc\bullet}(t) = \sum_{i \in S_{hc}} M_{hi}(t)$. Finally, (3.55) is easily re-written as

$$\begin{aligned} C_h^{1/2}\Big(\hat{\Lambda}_{0h}(t,\hat{\beta}) - \Lambda_{0h}(t)\Big) = {} & C_h^{1/2} \sum_{c=1}^{C_h} \psi_{hc}(t) \\ & - C_h^{1/2} g_h(\beta_0,t)' \frac{\Sigma_\beta^{-1}}{C_\bullet} \left(\sum_{h' \neq h} \sum_{c=1}^{C_h} U_{hc}(\beta_0) \right) + o_P(1) \ , \end{aligned} \tag{3.56}$$

where $\psi_{hc}(t)$ is given by (3.36). Note that the $\psi_{hc}(t)$'s are independent, but not identically distributed, terms.

The only random terms in the $\psi_{hc}(t)$'s are the $M_{hi}(t)$'s for $i \in S_{hc}$. Similarly, the only random terms in the second term on the right hand side of (3.56) are also the $M_{hi}(t)$'s but for $i \in \bigcup_{h' \neq h} S_{h'}$. Hence, the two terms on the right hand side of (3.56) have no random variables in common and, since by assumptions observations from different strata are independent, so are these two terms. In addition, it was already shown that $E(M_{hc\bullet}(t)) = 0$ and that $E(U_{hc}(\beta_0)) = 0$; see (3.14) and the discussion following (3.18), respectively. Therefore, both terms on the right hand side of (3.56) have mean zero. Hence, from (3.14) and the central limit theorem, $C_h^{1/2}\big(\hat{\Lambda}_{0h}(t,\hat{\beta}) - \Lambda_{0h}(t)\big)$ is asymptotically Gaussian with mean zero at each $t \in [0, \tau)$. Furthermore, by the functional central limit theorem (see Pollard (1990), page 53) the process

$$C_h^{1/2}\Big(\hat{\Lambda}_{0h}(t, \hat{\beta}) - \Lambda_{0h}(t)\Big) \qquad 0 \leq t < \tau$$

converges weakly to a zero-mean Gaussian process with covariance function $\sigma(t, s)$, given by (3.35).

The proof of the consistency of $\hat{\sigma}(t, s)$ is based on the consistency of $\mathcal{I}(\hat{\beta})$, the consistency of $G_h(\hat{\beta}, t)$ (see (3.50) and (3.51)), and on ideas similar to the ones used in dealing with (3.43)–(3.54); it is thus omitted. For further explanations see Lin et al. (2000), appendix A.3 and A.4. QED

It is worth noting some of the differences between theorem 3.4 and theorem 3 of Spiekerman & Lin (1998). Even though both have to account for clustering and are based on the stratified Cox model, there are fundamental differences between their proofs. In Spiekerman & Lin (1998), clusters are all of the same size and all observations experience the h different baseline hazard functions. This allows the proof of Spiekerman & Lin's (1998) theorem 3 to be based on the martingale central limit theorem; see section 1.3.4. As discussed earlier, this is not the case in section 3.2, and the functional central limit theorem must be used to prove theorem 3.4. A consequence of these differences is that $\hat{\sigma}(t, s)$, given by (3.38), has an extra term than $\hat{\xi}_{jk}(t, s)$, given on page 1168 of Spiekerman & Lin (1998). This extra term constitutes a direct contribution from the observations of the other $h - 1$ strata.

Both the proof of theorem 3.4 and the proof given in appendix A.4 of Lin et al. (2000) are based on the functional central limit theorem. However, as it allows for H different baseline functions, theorem 3.4 is a generalization of Lin et al.'s (2000). Hence, as it was the case with Spiekerman & Lin's (1998) $\hat{\xi}_{jk}(t, s)$, $\hat{\sigma}(t, s)$ has an extra term than $\hat{\xi}(t, s)$, given on page 716 of Lin et al. (2000). These differences and the reasons behind them will be discussed further in section 3.4.3.

3.2.4 Integrated hazard and survival functions

The integrated hazard function at time t for the i^{th} individual of the h^{th} stratum with fixed covariates x_{hi} can be estimated by

$$\hat{\Lambda}_{hi}(t) = \exp\{\hat{\beta}' x_{hi}\}\, \hat{\Lambda}_{0h}(t, \hat{\beta}) \ . \tag{3.57}$$

Analytical derivation of a formula for $\text{Var}\big(\hat{\Lambda}_{hi}(t)\big)$ that accounts for clustering is complex. First, a formula for the covariance between $\hat{\beta}$ and $\hat{\Lambda}_{0h}(t, \hat{\beta})$ must be derived. It is then possible, using the delta-method, to derive the formula in question.

A much simpler approach is to center all the covariates at x_{hi}, then $\hat{\Lambda}_{0h}(t, \hat{\beta})$ corresponds to $\hat{\Lambda}_{hi}(t)$, and $\hat{\sigma}_{\Lambda}(t, t)$, given by (3.38), can be used to estimate $\text{Var}\big(\hat{\Lambda}_{hi}(t)\big)$. To this end, re-run the analysis after replacing $(x_{11}, \ldots, x_{Hn_H})$ by $(x_{11} - x_{hi}, \ldots, x_{Hn_H} - x_{hi})$ in the dataset. This approach is also suggested by Lin et al. (2000).

Building on (3.57), the survival function at time t for the i^{th} individual of the h^{th} stratum with fixed covariates x_{hi} can be estimated by

$$\hat{S}_h(t \mid x_{hi}) = \exp\Big\{- \exp\{\hat{\beta}' x_{hi}\}\, \hat{\Lambda}_{0h}(t, \hat{\beta})\Big\} \ . \tag{3.58}$$

Using the delta-method and the idea of re-running the analysis after centering the covariates at x_{hi}, an uniformly consistent estimator of $\text{Var}\big(\hat{S}_h(t \mid x_{hi})\big)$ is

$$\frac{1}{C_h}\, \hat{\sigma}_{\Lambda}(t, t)\, \big(S_h(t \mid x_{hi})\big)^2 \ , \tag{3.59}$$

where $\hat{\sigma}_{\Lambda}(t, t)$ is given by (3.38).

3.3 Uninformative sampling designs

This section puts the methods proposed in section 3.2 in the context of analytical inference from longitudinal surveys. It is also meant as a continuation of sections 2.5–2.7. To this end, the notion of uninformative sampling designs is first defined and exemplified.

The notion of ignorable sampling designs, a concept closely related to uninformative, is discussed in section 3.3.1. The role of sampling weights when carrying out analytical inference from longitudinal surveys and, more generally, from complex survey data has been part of a long standing debate between statisticians. In sections 3.3.2, we hope to clarify the role and the use of sampling weights when carrying out analytical inference from longitudinal surveys.

The notion of uninformative sampling designs provides an important insight into the role of sampling weights when carrying out analytical inference from complex survey data. There are different variations of the definition of uninformative sampling design in the literature. Our definition 3.1 is based on the ones given by Binder & Roberts (2001) and by Pfeffermann (1993), section 3. A few minor modifications were made to make definition 3.1 more relevant to the particular case of the methods proposed in this chapter. First, let y_i denote the *response variable*, x_i denote the *vector of covariates* and z_i denote *design variables* or *design-related factors*, such as stratum and cluster information, for the i^{th} element of the population $U = \{1, \ldots, N\}$. The inclusion probability for the i^{th} individual, that is the probability that individual $i \in U$ is included in the sample $S \subseteq U$, is defined as

$$\pi_i(z_i, x_i, y_i) = \Pr\{R_i = 1 \mid z_i, x_i, y_i\} \ , \tag{3.60}$$

where $R_i = I(i \in S)$. The definition of an uninformative sampling design, given below, is based on these notions.

Definition 3.1 (Uninformative sampling design) A sampling design $p(S|Z)$ is uninformative if

$$\begin{gathered} \pi_i(z_i, x_i, y_i) = \pi_i(z_i, x_i) \quad \forall\, i \\ \text{and if} \\ \Pr\{y_i \mid x_i, R_i = 1\} = f(y_i \mid x_i) \quad \forall\, i \ ; \end{gathered} \tag{3.61}$$

the last condition is satisfied if Y_i and Z_i are independent given x_i. □

Obviously the structure of the population must be taken into account when carrying out inference. However, when the sampling design is uninformative, this does not need to be achieved via the randomization distribution $p(S|Z)$. Instead, the structure can be directly included within the probability model itself. Note that this will likely imply that x_i and z_i will share some components. Pfeffermann & Smith (1985) made similar comments when studying regression models.

The choice of the word uninformative is rather unfortunate. The assertion is not that the sampling design carries no useful information, but rather that, given the known values of the z_i's, the sampling design does not carry extra information. In other words, model-based or design-based analyses have access to all the information about stratification, clustering, etc., and should reflect the underlying super-population structure in the same way. Skinner et al. (1989), page 146, give an interesting and enlightening example of uninformative sampling design; other examples are given in chapter 4 and in Korn & Graubard (1999), section 4.3. A consequence of (3.61) is that analytical inference can be based on unweighted likelihood, score or estimating functions like

$$U(\theta) = \sum_{i \in S} \frac{\partial}{\partial \theta} \log f(y_i \mid x_i; \theta) \; . \tag{3.62}$$

Under mild conditions, solving $U(\theta) = 0$ provides a consistent estimator $\hat{\theta}$ of θ. In the case of the stratified Cox model given by (3.2), (3.62) becomes (3.7) and solving $U(\beta) = 0$ yields $\hat{\beta}$, which was shown to be consistent in theorem 3.1. Moreover, when the sampling design is uninformative, both weighted and unweighted procedures are unbiased and should yield similar point estimates (e.g., duration of jobless spells in section 4.2).

When (3.61) does not hold then the sampling design $p(S|Z)$ is said to be *informative*. In addition, if the π_i's depend on the y_i's then the sampling design is said to be *response-selective* or *response-dependent*. In either cases, (3.62) is no longer an unbiased estimating function and solving $U(\theta) = 0$ yields inconsistent estimates, with the exception of ignorable sampling designs described in section 3.3.1. The usefulness of $f(y \mid x; \theta)$ for analytic inference about individuals processes may be questionable in those circumstances. Nevertheless, pseudo-likelihood methods that take into account the inclusion probabilities can be used to make inference on θ. This is done by solving the following weighted *pseudo-score* function instead of (3.62)

$$U_W(\theta) = \sum_{i \in S} \frac{1}{\pi_i(z_i, x_i, y_i)} \frac{\partial}{\partial \theta} \log f(y_i \mid x_i; \theta) \ , \tag{3.63}$$

or, more generally,

$$U_{W'}(\theta) = \sum_{i \in S} w_{iS} \frac{\partial}{\partial \theta} \log f(y_i \mid x_i; \theta) \ , \tag{3.64}$$

where w_{iS}'s are such that

$$E\left(\sum_{i \in S} w_{iS} a_i \right) \approx \sum_{i \in U} a_i \quad \forall\, a = (a_1, \ldots, a_N)' \ .$$

The maximum pseudo-likelihood approach given by (3.63) and (3.64) is motivated by estimating the population parameter θ_U, where θ_U is the solution to $U(\theta) = 0$ with

$$U(\theta) = \sum_{i \in U} \frac{\partial}{\partial \theta} \log f(y_i \mid x_i; \theta) \ ; \tag{3.65}$$

see Binder (1983) for further information on this topic.

3.3.1 Ignorable sampling designs

It is possible to relax the assumption of uninformative sampling design, by requiring instead that the sampling design be ignorable, a weaker condition. The two notions are closely related, and uninformative implies ignorable. Let $\mathbf{z} = z_1, \ldots, z_N$, $\mathbf{y} = y_1, \ldots, y_N$ and $\mathbf{R} = R_1, \ldots, R_N$; then, the conditional joint distribution function of $\{y_i; i \in S\}$ is defined by

$$L(\theta) = f(y_i, \{i \in S\} \mid \{x_i; i \in S\}, \mathbf{z}) \ . \tag{3.66}$$

Another likelihood approach that includes the sampling design $p(S|Z)$ is based on the *pseudo-likelihood* function, which is defined as

$$L_{\mathcal{P}}(\theta) = \Pr\{\mathbf{R} \mid \mathbf{y}, \mathbf{z}\}\ f(y_i, \{i \in S\} \mid \{x_i; i \in S\}, \mathbf{z}, \mathbf{R}) \ . \tag{3.67}$$

In the spirit of Pfeffermann (1993), section 3, and based on (3.66) and (3.67) ignorable sampling design is defined below.

Definition 3.2 (Ignorable sampling design) A sampling design $p(S|Z)$ is ignorable if $\pi_i(z_i, x_i, y_i) = \pi_i(z_i, x_i)\ \forall\ i$ and if inference based on (3.66) is the same as inference based on (3.67). □

The notion of ignorability refers to the information provided by the sampling design $p(S|Z)$ in addition to what is already provided by the design variables z_i's. Sugden & Smith (1984) explore the conditions under which a sampling design, depending only on the z_i's, is ignorable given only partial information on the design in question. The ideas behind ignorable sampling designs are closely related to the notion of missing at random and observed at random introduced by Rubin (1976) for the ignorability of the process causing missing values. Hence, when the sampling design is ignorable (3.62) is an unbiased estimating function. Nevertheless, the fact that $\Pr\{y_i \mid x_i, R_i = 1\} = f(y_i \mid x_i)$ does not necessarily hold when the sampling design is ignorable, implies that making analytical inference is not as "meaningful" as when the sampling design is uninformative. See Binder & Roberts (2001) for further discussion on uninformative and ignorable sampling designs, and differences between the two.

It is important not to confuse the notions of uninformative and ignorable sampling designs with the ones of independent and noninformative censoring (see section 1.2.4). However, if censoring depends on the z_i's in such a way that it is not independent of T_i, given x_i, then neither (3.62) nor (3.63) are unbiased estimating functions. In that case, it would be necessary to use time-varying weights in (3.63) or (3.64) in order to have $E(U_W(\theta)) = 0$, a fact that has been generally overlooked; see Lawless (2003b), section 6.

Testing the ignorability conditions

In theory, when the design variables z_i's are known for all the population units and when all the design features are known, it is possible to verify if the sampling design is ignorable using definition 3.2. Similarly, it is possible to verify if all the conditions for the sampling design to be uninformative are fulfilled using definition 3.1. In many practical cases, however, only limited knowledge about the actual sampling design is available. In addition, not all the relevant design variables are known for all the populations units, or incorporating all of them in the model is either not feasible or suitable. Not incorporating all the design variables in the model does not necessarily imply that (3.62) is a biased estimating function. As indicated by Sugden & Smith (1984), incorporating only partial design information in the model can be sufficient to obtain an ignorable sampling design. Therefore, being able to test if a sampling design is ignorable is valuable.

Such a test is given by Pfeffermann (1993), section 4.2. It is based on the principal that $\hat{\theta}_W$ obtained by solving $U_W(\theta) = 0$, where $U_W(\theta)$ is given by (3.63), is always a consistent estimator of θ_U (i.e., regardless if the sampling design is ignorable or not). If the model holds in the population U, then $\lim_{n\to\infty,\, N\to\infty} \hat{\theta}_W = \theta$ in probability. However, $\hat{\theta}$ obtained by solving $U(\theta) = 0$, where $U(\theta)$ is given by (3.62), is only consistent when the sampling design is ignorable. Hence, the two estimators converge to θ under ignorable sampling design. Unfortunately, the two estimators will converge to different limits when the model in the population U is wrongly specified. Thus, convergence to the same limit is sufficient, but not necessary for ignorability of the design. If $\hat{\theta}_W$ is asymptotically normal, the test statistic given by Pfeffermann (1993) and its asymptotic distribution are

$$(\hat{\theta}_W - \hat{\theta}) \left(\widehat{V}_W(\hat{\theta}_W) - \widehat{V}(\hat{\theta})\right)^{-1} (\hat{\theta}_W - \hat{\theta}) \sim \chi^2_p \,, \tag{3.68}$$

where $p = \dim(\theta)$ and

$$\widehat{V}_W(\hat{\theta}_W) = \mathcal{I}_W^{-1}(\hat{\theta}_W)\, \widehat{V}\big(U_W(\hat{\theta})\big)\, \mathcal{I}_W^{-1}(\hat{\theta}_W) \,,$$

with $\mathcal{I}_W(\theta) = -\partial U_W(\theta)/\partial\theta$. See Binder (1983) for the four main assumptions under which $\widehat{V}(\hat{\theta}_W)$ is consistent, as well as further explanations on variance estimation of $\hat{\theta}_W$. In Pfeffermann (1993), $\widehat{V}(\widehat{\theta}) = \mathcal{I}^{-1}(\hat{\theta})$; however, robust variance estimators like the ones proposed by Lawless & Boudreau (2002) or Lawless (2003b) could be used instead without affecting the limiting distribution. Lawless's (2003b) robust variance estimators account for clustering and stratification, and is in the same spirit as $\widehat{V}_R(\hat{\beta})$ given by (3.19). In addition to the sufficient but not necessary limitation mentioned previously, (3.68) also suffers from lack of power. This is due to the fact that weighted analyses tend to increase the variance of the estimated parameters; this is discussed further in section 3.3.2.

For these reasons (3.68) is not widely used to test if the sampling design is ignorable or not. Since uninformative implies ignorable, it also gives some information on whether or not the sampling design is uninformative. Simple residual plots against known design variables or the inclusion probabilities are other useful tools in accessing if the sampling design is uninformative; see Pfeffermann & Smith (1985) for an example using data from an Israeli survey on the productivity of grapefruits in 1979.

Korn & Graubard (1999), section 4.4, also give a test to help in choosing between weighted or unweighted analyses. This test does not look at the ignorability of the sampling design, but rather at the inefficiency of using sample weights when unnecessary for estimating θ. In other words, if the price of using weights is too high in terms of variance, then use unweighted analyses, otherwise use weighted analyses. Unfortunately, it is easy to come with an example where the sampling design is informative (or not ignorable) and where weighted estimates have much larger variances that unweighted ones. In such a case, Korn & Graubard's (1999) test would wrongly indicate that an unweighted analysis is preferable. Choosing between weighted or unweighted analyses should be dictated by the sampling design being uninformative and/or ignorable and not by the cost in efficiency of weighted analyses; see Hoem (1989), Pfeffermann (1993) or Lawless (2003b). Although it is helpful, one has to be careful not to read too much into Korn & Graubard's (1999) test.

Part of the reasons behind definitions 3.1 and 3.2 was to facilitate the discussion in section 3.3.2 and make it more meaningful.

3.3.2 Weighted versus unweighted analyses

Sampling weights have been used for many decades and are widely accepted by the statistical community when carrying out descriptive inference from survey data; that is, inference about known functions of a finite population parameters (see section 2.7). In standard survey practice or descriptive inference, weights are mainly used to compensate for unequal selection probabilities. Unequal selection may arise because it is beneficial or cost-effective to over-sample particular subgroups of the population or sample subgroups at different rates. For example, particularly influential observations may have selection probabilities close (or equal) to 1, making sure that they are generally (or always) included in the selected sample. Other reasons include accounting for non-response and the use of clusters when a list that identifies each and every element of the population is unavailable, or that producing such a list is too costly or impractical.

In contrast, there has been a long standing debate in the statistical community on the use and role of sampling weights when carrying out analytical inference about model parameters from survey data. At one end are statisticians like Hoem and Fienberg, for whom sampling weights are at best irrelevant to model-based analytical inference, except when the sampling design is response-selective[1]; see Hoem (1989) and Fienberg (1989). At the other end, Binder, Folsom and Kalton are strong advocates of the use of sampling weights. Their main argument in favor of using sampling weights are that they yield estimators that are more robust at a low cost in efficiency; see Binder (1983), Kalton (1989) and Folsom et al. (1989). However, there is a consensus that, for model-based inference, weighted and unweighted results should be similar. An appreciable difference is an indication that the model is false or misspecified and that a modified or better model should be sought; see Kalton (1989), page 581. This last point is also illustrated by the various stratified Cox PH models fitted in sections 4.1 and 4.2. Before describing and justifying the point of view chosen in this thesis, I want to make clear that not all of these authors have the same position on all relevant issues and that theirs represent only a small sample of interesting and sometimes provocative opinions on this topic. I also recognize that opinions, including mine, develop over time. In addition, I do not have the pretension that the following arguments will end the debate on why and when

[1] Referred to as outcome-based sampling in their papers

to use sampling weights.

The point of view taken in this thesis is between the two extremes mentioned in the previous paragraph and in line with the notions of uninformative (or ignorable) sampling design defined in section 3.3. Accordingly, use unweighted analyses when the goal is analytical inference for parameters of a super-population or a probabilistic model and when the sampling design is uninformative or ignorable; otherwise, use weighted analyses. In other words, use (3.62) in the first case and (3.63) or (3.64) in the second. Note that the importance of analytical inference in the context of longitudinal surveys was already discussed in section 2.7, and that this recommendation on when to use sampling weights comes with the following warning. When doing weighted analyses one has to keep in mind that results are only meaningful for that particular finite population from which the sample was drawn and at that given time; that is, they have descriptive values only. They do not extend to another population or to the same population at a future time point. Thus, the usefulness of these results is extremely limited when the goal is to understand the history processes of individuals or causal mechanisms. For example, define

$$S_z(t) = \Pr\{T \geq t \mid Z = z\} \ , \tag{3.69}$$

where z is the (discrete) design variable, and

$$\bar{S}(t) = \Pr\{T \geq t\} = \sum_z \Pr\{Z = z\} S_z(t) \ . \tag{3.70}$$

The descriptive meaning and usefulness of (3.70) are clear. Hence, $\bar{S}(t)$ is an estimate of the average duration time for the combined individuals of a given surveyed population at a particular time. This estimate accounts for the different values of z in the proportions that they appear in the surveyed population in question at that given time. However, as the survival times T_i's are related to the z_i's and possibly to other covariates, the use and meaning of (3.70) beyond descriptive purposes are both limited and not clear. In other words, it is hard to justify how $\bar{S}(t)$ computed using data from a given surveyed population at a particular time could also represent or explain what is happening in another population, or in the same population but at another time. Note that the point we are making is not that quantities like $\bar{S}(t)$ are not of interest, but rather that models are easier to generalize.

The reasoning behind our position on the use of sampling weights is based on:

1. Estimators from unweighted model-based procedures are unbiased when the sampling design is uninformative (or ignorable), see the implications of definition 3.1 on (3.62);

2. These estimators are, from the theory of maximum likelihood, consistent and asymptotically efficient when the sampling design is uninformative (or ignorable).

Hence, using sampling weights implies a loss in efficiency or increased estimator variances. In addition to what was mentioned in section 2.6 on the difficulties of defining "proper" sampling weights in the context of longitudinal surveys, there is no clear principle in the choice of design consistent estimators[2] in the literature. This means that it is possible to have more than one weighted estimator for a given parameter and no clear criteria to decide which one to choose. This point is discussed in further detail in Little (1989) and Pfeffermann (1993), section 7. In addition, if one believes sufficiently in the model to define the parameter θ_U as the solution of $U(\theta) = 0$, where $U(\theta)$ is given by (3.65), then one should be even more confident that the appropriate score function is (3.62), and there should be no need for (3.63) or (3.64). This last point was also raised by Thompson (1997), section 6.2.

Among the arguments raised by those in favor of using sampling weights, when carrying analytical inference and when the sampling design is uninformative or ignorable, are that they provide estimators that are more robust at a cost in efficiency that is acceptable. However, the size of that loss of efficiency is not mentioned nor is anything said about how much is gained in robustness. Looking at the examples of section 4.2, the lost

[2] The sample statistics $\tilde{\theta}_S$ is said to be a design consistent estimator of θ_U if $\lim_{n\to\infty,\, N\to\infty} \tilde{\theta}_S - \theta_U = 0$ in probability, where n is the sample size and N is the population size.

in efficiency can be substantial. Some authors refer to the previous as protection against "model misspecification", but the real issue is one of informative versus uninformative sampling designs. As mentioned by Hoem (1989), sampling weights were not devised to protect against such misspecification in model-based analysis and there exists no proof that they can serve this function. In addition, no strong assumption about the model is made when non-parametric or semi-parametric methods are used (although this could be open for discussion in the case of semi-parametric methods). For example, Lawless (2003b) describes a method in the same spirit as the one proposed in section 3.2.2, but for the variance of the Kaplan-Meier estimator. His method, which accounts for clustering but not for stratification, shares similarities with the one by Williams (1995). If the assumption is made that the sampling design is uninformative (or ignorable) when it is not, or the information given by the z_i's is overlooked, then weighted procedures will still yield unbiased estimators while unweighted procedures will not. In this context, sampling weights have a function as a guard against "model misspecification". Unfortunately, this is achieved at the expense of the ability to make more general inference from the data, as discussed and illustrated by (3.70). In any case, model misspecification and informative sampling are distinct issues and should not be confused.

The importance of analytical inference in longitudinal studies and in sample-based event history analysis, which is generally based on a probabilistic model, was already emphasized and discussed in section 2.7 (e.g., see the comments and explanations following (2.1) and (2.2)). The same point was also made by Pfeffermann & Smith (1985) in the context of linear regression. In their article, they mentioned that they favour an analytic approach to regression analysis as they find it difficult to justify the estimation of β_U, the descriptive parameter[3], without relying on a well defined model and on β_U being close to β, the analytical parameter. This is motivated by their belief that inference about relationships is usually meant to refer to populations more general than the fixed finite

[3] Pfeffermann & Smith (1985) defines $\beta_U = (X'X)^{-1}X'Y$, where X is a $N \times p$ matrix and Y is a $N \times 1$ vector, with $Y_i = \beta' X_i + \varepsilon_i$ for $i \in U = \{1, \ldots, N\}$.

population U, which existed at the time the sample was drawn. As mentioned, the goal of carrying analytical inference is to be able to reach conclusions that go beyond the finite population under study. This implies using a probabilistic model, and that the model in question holds regardless of $S \subseteq U$ and z_i's. In other words, (3.61) must hold and the sampling design is uninformative. Again, it is possible to relax this condition by requiring the sampling design to be ignorable, which implies that the same inference, and thus conclusions are reached with both unweighted score and weighted pseudo-score functions, see definition 3.2.

Sugden & Smith (1984) discuss the conditions under which a sampling design is ignorable. Some of these conditions could be adapted to the case of uninformative sampling design. However, a simpler strategy was chosen in this thesis to make sampling designs uninformative. The idea is to incorporate the z_i's and other information about the sampling design into the probabilistic model as covariates or by other means to ensure that (3.61) holds. As (3.2) is a semi-parametric model and fairly flexible, the necessary information could be incorporated without major problem. Hence, different strata had different baseline hazard functions, clustering was handled by the proposed robust variance estimators of section 3.2 and the other design variables were included as covariates. Although not expressed in the same terms, Patterson et al. (2002) also made the same connections between unweighted analyses, uninformative sampling and analytical inference. Their actual arguments are given on page 727 of their article and are based on the work of Korn & Graubard (1999).

As mentioned earlier, the above discussion does not pretend to end the debate on the use of sampling weights when carrying analytical inference. The different authors in the book edited by Kasprzyk et al. (1989) provide a wide range of pros and cons for both weighted and unweighted analyses. Other references include Skinner et al. (1989), Pfeffermann (1993), Korn & Graubard (1999), Patterson et al. (2002) and Lawless (2003b).

Weighted versus unweighted as diagnostic tools

As was mentioned in section 3.3.1, weighted and unweighted analyses should agree when the sampling design is uninformative or ignorable. If they do not, it is a good indication of one or more of the following:

1. The sampling design is informative or non-ignorable;
2. The probabilistic model is misspecified or wrong;
3. Part or all of the design information given by the z_i's is not accounted for by the model. This is especially useful in event history analysis where diagnostic tools are not as extensive as for other statistical methods.

This last point is illustrated by the following example. Assume that the length of unemployment spells for an imaginary population of workers is of interest. The population is divided in two strata (high and low income), and a simple random sample is drawn from each stratum. The sampling design is such that a larger sampling fraction is used for people with high income than for people with low income. Also, people with high income have shorter spells of unemployment than people with low income. Assuming that the analyst does not have access to income information for each individual, the sampling design is then informative. Nevertheless, this is unknown or ignored by the analyst, and only one survival curve is fitted to the entire population using an unweighted Kaplan-Meier estimator. Obviously, this estimator is meaningless as it depends on the choice of the sampling fractions. A weighted Kaplan-Meier estimate will give more weight to high-income people than to low-income people, and will not agree with the unweighted one. However, the weighted Kaplan-Meier estimator will provide more meaningful results, in the same sense as (3.70). That is, the weighted Kaplan-Meier gives a valuable estimate of the length of unemployment spells for the population of high and low income people combined, taking into account the proportions of high and low income people in the surveyed population at that given time. However, the weighted estimate is not of much use for analytic inference. In the present example, the unweighted estimation could easily be corrected by fitting two separate survival curves for high and low income people, assuming that information is available somehow. Another easily corrected example is given by Kalton (1989). On the other hand, it is not difficult to come out with much more complex examples.

This procedure of comparing weighted and unweighted estimators shares some ideas with the method described by Fitzmaurice et al. (1997) for detecting over-dispersion in large scale surveys. They suggest grouping the observations into C independent clusters. These groups should be constructed in such a way that there is no substantive interpretation about differences in survival functions across clusters. Next, compute $\hat{\theta}_c$, the estimate of θ obtained by using only the observation from the c^{th} cluster, for $c = 1, \ldots, C$. Combined the values of $\hat{\theta}_c$'s in the same fashion as to get a jackknife variance estimator; this results in the following variance estimator,

$$\widehat{V}_C(\hat{\theta}) = \frac{1}{C(C-1)} \sum_{c=1}^{C} (\hat{\theta}_c - \bar{\theta}_C)^2 \ , \tag{3.71}$$

where

$$\bar{\theta}_C = \frac{1}{C} \sum_{c=1}^{C} \hat{\theta}_c \ .$$

It is then easy to compare $\widehat{V}_C(\hat{\theta})$ to $\mathcal{I}^{-1}(\hat{\theta})$, the variance of $\hat{\theta}$ obtained under the assumption that all observations are independent. If the ratio is larger than 1, there is evidence of over-dispersion in the data, and $\widehat{V}_C(\hat{\theta})$ is preferable. As mentioned in section 3 of their paper, differences between $\widehat{V}_C(\hat{\theta})$, an unweighted variance estimator, and $\widehat{V}_W(\hat{\theta})$, a weighted variance estimator given in (3.68), indicate extra variation not accounted for by the sampling design. This extra variation or over-dispersion is often the result of the omission of important predictors in the probabilistic model. These last sentences rejoin one of our comments on the implications of differences between weighted and unweighted analyses; see point 2 on page 84.

3.4 Remarks on proposed methods

This section concludes chapter 3 by discussing some software issues regarding the computation of the estimators given in section 3.2 (section 3.4.1), deriving weighted versions of these estimators (section 3.4.2), making a few comments on methods similar to the ones proposed in this chapter (section 3.4.3), and illustrating how these proposed methods can be applied to other types of studies (section 3.4.4).

3.4.1 Computer issues

Marginal proportional hazard models discussed in this chapter are easily implemented using SAS or S-Plus. Recent releases of these software[4] make it easy to get the variance estimates given in section 3.2.2. In SAS, the PHREG procedure with its strata statement will fit model (3.2) and the keyword ressco, in the output statement, will yield the score residuals for every individual. The sums of these residuals for each cluster can easily be computed as the id statement adds an extra variable to the outputted dataset indicating to which cluster each individual belongs. It is then easy to compute $\widehat{V}_R(\hat{\beta})$ given by theorem 3.3. Fitting these models and getting $\widehat{V}_R(\hat{\beta})$ is even simpler in S-Plus. The coxph function, in conjunction with the strata and cluster statements in the formula argument, yields the values of $\hat{\beta}$ and of $\widehat{V}_R(\hat{\beta})$. Moreover, the sum of the score residuals for each cluster can be obtained using the residuals function, with the type='score' and collapse arguments.

The fact that S-Plus gives $\widehat{V}_R(\hat{\beta})$ is mainly due to programmers who made the coxph function very flexible. Even though there was no theoretical results on stratified Cox's models with clusters of various sizes before Boudreau & Lawless (2001), coxph was designed to handle these models.

Point estimates for $\hat{\Lambda}_{0h}(t, \hat{\beta})$, $\hat{\Lambda}_{hi}(t)$ and $\hat{S}_h(t \mid x_{hi})$ (see (3.34), (3.57) and (3.58)) are

[4] S-Plus for Windows version 4.0 or higher, S-Plus for Unix/SunOS version 5.0 or higher and SAS release 6.07 or higher for both Windows and Unix/SunOS.

easily obtained from SAS or S-Plus. In SAS, this requires using the baseline statement of the PHREG procedure. In S-Plus, the same is accomplished by using the survfit function with the corresponding fitted coxph object as its first argument. Note that in that case the default is to use the Fleming-Harrington estimate of the survival curve, which is (3.58). Unfortunately, the variance estimator introduced in section 3.2.3 is not implemented in either software and extra coding is required.

3.4.2 Weighted versions of the estimators of section 3.2

Weighted versions of the proposed variance estimators given in section 3.2 are easily derived. First, redefined $S_h^{(r)}(\beta, t)$, for $r = \{0, 1\}$, given by (3.4) and (3.5) as

$$S_h^{(0)}(\beta, t) = \frac{1}{n_h} \sum_{i=1}^{n_h} w_{hi}\, \gamma_{hi}(t)\, \exp\{\beta' x_{hi}(t)\} \tag{3.72}$$

and

$$S_h^{(1)}(\beta, t) = \frac{1}{n_h} \sum_{i=1}^{n_h} w_{hi}\, \gamma_{hi}(t)\, x_{hi}(t)\, \exp\{\beta' x_{hi}(t)\}\ , \tag{3.73}$$

for $i = 1, \ldots, n_h$ and $h = 1, \ldots, H$.

Therefore, the weighted version of (3.7) is given by

$$U_W(\beta) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi}\, \delta_{hi} \left(x_{hi}(t_{hi}) - \frac{S_h^{(1)}(\beta, t_{hi})}{S_h^{(0)}(\beta, t_{hi})} \right)\ , \tag{3.74}$$

and the estimator $\hat{\beta}_W$ is defined as the solution of $U_W(\beta) = 0$.

Using the same arguments and reasoning as the ones used in the derivation of (3.19), a weighted robust variance estimator for $\hat{\beta}_W$ is given by,

$$\widehat{V}_{R_W}(\hat{\beta}_W) = \mathcal{I}_W^{-1}(\hat{\beta}_W) \left(\sum_{h=1}^{H} \sum_{c=1}^{C_h} (w_{hc\bullet})^2\, \hat{U}_{hc}(\hat{\beta}_W)\, \hat{U}_{hc}(\hat{\beta}_W)' \right) \mathcal{I}_W^{-1}(\hat{\beta}_W)\ , \tag{3.75}$$

where $\mathcal{I}_W(\beta) = -\partial U_W(\beta)/\partial \beta'$, $w_{hc\bullet} = \sum_{i \in S_{hc}} w_{hi}$ and $\hat{U}_{hc}(\hat{\beta})$ is given by (3.20).

The weighted Breslow-Aalen estimator for the baseline function is given by

$$\hat{\Lambda}_{0h_W}(t,\hat{\beta}) = \int_0^t \frac{w_{h\bullet}\, J_h(u)}{n_h S_h^{(0)}(\hat{\beta},u)}\, dN_{h\bullet}(u) \quad \text{for } h = 1,\ldots,H \ , \tag{3.76}$$

where $N_{h\bullet}(t)$ and $J_h(t)$ were defined as part of (3.34) and $w_{h\bullet} = \sum_{i=1}^{n_h} w_{hi}$. The derivation of the weighted robust variance estimator for (3.76) is based on the same reasoning as the ones used in the derivation of (3.34). Hence, $\text{Var}\big(C_h^{1/2}\hat{\Lambda}_{0h_W}(t,\hat{\beta})\big)$ is estimated consistently and uniformly in t by

$$\begin{aligned} \hat{\sigma}_{\Lambda_W}(t) = C_h \sum_{c=1}^{C_h} \hat{\psi}_{hc_W}(t)^2 &+ C_h\, G_h(\hat{\beta}_W,t)'\,\mathcal{I}_W^{-1}(\hat{\beta}_W) \\ &\times \left(\sum_{h'\neq h} \sum_{c=1}^{C_{h'}} (w_{hc\bullet})^2\, \hat{U}_{h'c}(\hat{\beta}_W)\, \hat{U}_{h'c}(\hat{\beta}_W)' \right) \mathcal{I}_W^{-1}(\hat{\beta}_W)\, G_h(\hat{\beta}_W,t) \ , \end{aligned} \tag{3.77}$$

where $G_h(\beta,t)$ is given by (3.41), and

$$\hat{\psi}_{hc_W}(t) = \sum_{i\in S_{hc}} \int_0^t \frac{1}{n_h S_h^{(0)}(\hat{\beta}_W,u)}\, d\hat{M}_{hc\bullet}(u) - w_{hc\bullet}\, G_h(\hat{\beta}_W,t)'\,\mathcal{I}_W^{-1}(\hat{\beta}_W)\,\hat{U}_{hc}(\hat{\beta}_W) \ ,$$

with $S_h^{(0)}(\beta,t)$ defined by (3.72).

Note that, more research is needed to study the properties of (3.75) and (3.77); see section 6.2. In addition, see the end of section 3.4.3 for comments on the similarities between these weighted estimators and Lin's (2000) super-population variance estimators.

3.4.3 Comments on similar methods

Lin & Wei (1989) were among the first authors to work on a similar problem. Their "robust" variance estimator for $\hat{\beta}$, which is in the same spirit as our $\widehat{V}_R(\hat{\beta})$ given in theorem 3.3, was developed to protect against model misspecification. Subsequent work branched into two general frameworks. The first one deals with clustered or recurrent failure times (e.g., Wei et al. (1989), Lee et al. (1992), Liang et al. (1993), Spiekerman & Lin (1998) and Lin et al. (2000)). The second one deals with model-based inference from complex survey data (e.g., Binder (1992) and Lin (2000)).

Wei et al. (1989) assumed that a cluster corresponds to an individual and that each individual is at risk of K types of failure, or to the same type of failure but for K successive occasions. Thus K failure or censoring times, most likely correlated, are recorded for each individual and, consequently, all clusters are of size K. The stratified Cox model they use is slightly different than (3.2) and allows for a stratum specific (or individual specific, in their case) β_0; that is, β_0 is replaced by β_{0h} in (3.2). Regardless of the differences between their setting and ours, most of the ideas used in the derivation of their "robust" variance estimator for $\hat{\beta}_{0h}$, denoted $\hat{Q}$, are the same ones than those used in the derivations of (3.19). However, the assumption that all the clusters are of the same size simplifies the derivation of $\hat{Q}$. The assumptions of section 3.2.1 are less restrictive and allow for clusters of various sizes. Also, any given individual can be a member of only one of the H strata, which correspond to the H baseline hazard functions.

The methods proposed by Lee et al. (1992) have in common with ours that cluster[5] sizes can vary and, as a result, their variance estimator (5.1) shares similarities with our $\widehat{V}_R(\hat{\beta})$. However, their model does not allow for different baseline functions and their proof requires assuming the existence of "ghost" individuals. Moreover, their "robust" variance estimator for the survivor function does not take into account clustering between individuals. Like the method proposed in section 3.2.2 and the one by Lee et al. (1992), Liang et al. (1993) allow for clusters of different sizes. Their proof relies on expressing their estimating equation as a U-statistic. This leads to a very different proof and to a biased variance estimator, which is then made unbiased (see (11) in their paper) by multiplying by a factor, which is function of the number of clusters. Spiekerman & Lin (1998) extended the work of Wei et al. (1989) to allow for more than one individual per cluster. However, they assumed that clusters are all of the same size, which allows them to express $U(\beta)$ as sums of i.i.d. random vectors and to use the multivariate central limit theorem to prove their theorem 1. In contrast, we only assumed that $U(\beta)$ can be expressed as a sum of independent random vectors. The problem discussed by Lin

[5] Strangely referred to as strata in their article.

et al. (2000) is different that ours. Their focus is on recurrent events, thus correlation arises from individuals experiencing multiple events and not from individuals sharing the same household. Also, they only allow for one baseline hazard function, compare to H is chapter 3. However, both the proofs of section 3.2 and of Lin et al. (2000), appendix A.1–A.5, rely on empirical process theory.

In summary, the methods proposed in section 3.2 combined the advantages of the methods introduced by Lin & Wei (1989), Wei et al. (1989), Lee et al. (1992), Liang et al. (1993), Spiekerman & Lin (1998) and Lin et al. (2000). That is, they allow for both clusters of various sizes and for different baseline hazard functions. In addition, variance estimators for $\hat{\Lambda}_{0h}(t, \hat{\beta})$, $\hat{\Lambda}_{hi}(t)$ and $\hat{S}_h(t \mid x_{hi})$ that account for intra-cluster dependence were derived in sections 3.2.3 and 3.2.4.

Binder (1992) extended the work by Lin & Wei (1989) to the context of survey data. Possible correlation between observations is accounted for by the design-based variance estimator of $\hat{\beta}$ he derived. Lin (2000) gives a formal justification for Binder's variance estimator, using the properties of the normalised Horvitz-Thompson estimator. He also generalized Binder's work to the context of super-population analytical inference. However, the additional term in the variance estimator that accounts for the extra variation due to the super-population inference assumes independence between all observations. This assumption may be violated if the clusters are not only the result of the complex sampling design, but are part of the structure of the super-population under study (e.g., the super-populations corresponding to the SLID and SIPP data of section 1.1 are composed of households of individuals). Weighted versions of the proposed variance estimators, which do not suffer from that problem, were derived in section 3.4.2. These weighted estimators and Lin's (2000) super-population variance estimators yield values that are very close for settings that we have examined; see tables 4.2, 4.4 and 4.7 as well as their discussion.

3.4.4 Applications to other studies

Although the methods introduced in this chapter are mainly designed for longitudinal studies based on complex survey designs, they can also be applied to other types of studies. For example, a cluster can consist of multiple observations on a given patient (e.g., measurements on both eyes, recurrent events, etc.) instead of individuals living in the same household as in section 1.1. Alternatively, a cluster can be two patients in a matched pair clinical trial or, more generally, matched patients in a case-control study.

Similarly, the subscript h in the stratified Cox model given by (3.2) could refer to male versus female or smoker versus non-smoker instead of the different strata of a population $U = \bigcup_{h=1}^{H} U_h$. Other examples include using different baselines for different histological types of malignant melanoma cancer or for different hospitals in a multi-center clinical trial study.

A common covariate effect (i.e., constant β_0) across strata has been assumed. One should of course be aware of the possibility for stratum-covariate interactions. The methods here are readily extended to incorporate this. In fact, by defining covariate vectors $x_{hi}(t)$ so that certain components are zeros in certain strata, the methodology proposed in chapter 3 can be apply to cases where some covariates have stratum-specific effects. Finally, in studies where the number of strata is large it may sometimes be desirable to combine certain strata for analytical purposes; this is easily handled by unweighted analyses.

In the next chapter, the methods proposed in section 3.2 are illustrated using the Survey of Income and Program Participation (SIPP) and the Canadian Survey of Labour and Income Dynamics (SLID).

Chapter 4

Examples

Regardless of the asymptotic properties of a statistical method, it is always of interest to see how it performs when applied to real life datasets. In this chapter, the methods proposed in chapter 3 are applied to the Survey of Income and Program Participation (SIPP) and the Survey of Labour and Income Dynamics (SLID). Section 4.1 focuses on analytical inference from the 1987 SIPP dataset, described in section 1.1.2. In particular, it is concerned with the effects of sex, age, race, education and individual income on the duration of spells on the food stamps program (see section 4.1.1) and of spells without health insurance (see section 4.1.2). Both weighted and unweighted versions of model (3.2) are fitted using a total of 1,816 uncensored and right-censored spells on the food stamps program. Similarly for spells without health insurance, analyses are based on a total of 7,656 spells. No left-truncated spell is included in the different analyses; see table 2.1 for the prevalence of left-truncation in SIPP. In section 4.2, the focus is on analytical inference from the 1993–1998 SLID dataset (i.e., the 1st panel), described in section 1.1.1. The focus is on the effects of sex, age, children, education, wage, employment insurance, etc. on the duration of jobless spells. Again, both weighted and unweighted versions of (3.2) are fitted using a total of 16,682 uncensored and right-censored spells. As in section 4.1, no left-truncated spell is included. Comparing weighted and unweighted analyses will allow us to explore some of the issues raised in section 3.3 in a real life context. Finally, another goal of chapter 4 is to compare the methods proposed in chapter 3 with the ones introduced by Binder (1992) and Lin (2000).

A great amount of time and care went into the examples presented in chapter 4. In addition, they are among the few SIPP and SLID analyses with purely analytical goals. Regardless of that, their main purpose was to illustrate the methods proposed in chapter 3 and caution is required when viewing them as comprehensive assessments of the data. References on SIPP data analyses include: Young (1989), Kalton & Miller (1991) and Kalton et al. (1992). With its 1st panel ending in 1998, SLID is a fairly new longitudinal survey compared to SIPP or PSID, and fewer comprehensive data analyses have been done so far. References on SLID include: Michaud & Webber (1994) and Cotton & Giles (1998).

4.1 SIPP analyses

The 1987 SIPP panel data was obtained through the Inter-University Consortium for Political and Social Research (ICPSR), a unit of the Institute for Social Research at the University of Michigan; see `http://www.icpsr.umich.edu`. The University of Waterloo is among the several hundred institutions member of the ICPSR. The dataset in question is very large, approximately 430 Meg's, and SAS was used to extract the relevant variables. The information on spell durations was recoded in a format suitable to most statistical software using C++. Since our procedures are based on continuous time variables, spell lengths were converted from a monthly discrete scale to a continuous or daily scale. This was done by converting a spell of m months into a spell of $(m-1) \times 30 + [u]$ days, where $U \sim \text{Unif}(0, 31)$ and $[u]$ is the largest integer smaller or equal to u. Recall that the main goal of section 4.1 is to illustrate the proposed methods and is not intended as a comprehensive data analysis. Nevertheless, it would be interesting to extend the methods proposed in chapter 3 to handle discrete time variable.

Stratified Cox proportional hazards models, as given by (3.2), were fitted to the duration (in days) of spells on the food stamps program and of spells without health insurance in sections 4.1.1 and 4.1.2, respectively. These models are composed of 46 strata, each having its own baseline hazard function $\lambda_{0h}(t)$, $h = 1, \ldots, 46$. These strata basically correspond to the 50 states, the District of Columbia and individuals who could not be classified in any of the previous (e.g., living outside the U.S. or having lived in more than one state). In addition, the following 9 states were joined into 3 groups: 1) Maine and Vermont; 2) Iowa, North Dakota and South Dakota; and 3) Alaska, Idaho, Montana and Wyoming. These states were grouped together because some of them have a fairly small population compared to others (e.g., Wyoming with an estimated total population of 493,423 as of July 1, 2001[1]). In addition, they are geographically close together, except for Alaska.

The following covariates were included in the exponential term of model (3.2):

- Sex: 1 — man (baseline) and 2 — woman; that is,

$$I(\text{Sex}) = \begin{cases} 1 & \text{if the individual is a woman} \\ 0 & \text{otherwise ;} \end{cases}$$

- Race: 1 — white (baseline), 2 — black, 3 — American Indian, Eskimo or Aleut and 4 — Asian or Pacific islander; that is,

$$I(\text{Black}) = \begin{cases} 1 & \text{if the individual is black} \\ 0 & \text{otherwise ,} \end{cases}$$

$$I(\text{Indian}) = \begin{cases} 1 & \text{if the individual is American Indian, Eskimo or Aleut} \\ 0 & \text{otherwise ,} \end{cases}$$

$$I(\text{Asian}) = \begin{cases} 1 & \text{if the individual is Asian or Pacific islander} \\ 0 & \text{otherwise ;} \end{cases}$$

- Age;

[1] Source: U.S. Census Bureau.

- Education or years of school attended;
- Total individual monthly income (in dollars).

These three later covariates are recorded at every four-month interval (barring missing values). To simplify the examples, the average of these recorded total monthly incomes, values for age and years of school attended was computed for each individuals, and these individual averages were used in fitting model (3.2).

The other two variables that are part of the analyses of sections 4.1.1 and 4.1.2 are the longitudinal sampling weights, for the weighted analyses, and the cluster or house ID. The former refers to the strictly positive longitudinal weights assigned to each member of a household selected at the start of the study in February 1987; that is 24,458 individuals (or 68%) of the sample. The 11,516 (or 32%) others who joined existing households afterward have sampling weights of zero, as discussed in section 1.1.2. The information on PSU's is, unfortunately, unavailable to people outside the U.S. Census Bureau; thus, they can not be included in our model. Part of the correlation, not accounted for by the covariates mentioned above, is due to individuals experiencing multiple spells and to the choice of following every member of selected households; see Citro et al. (1986) for the definition of "household" adopted by SIPP. This latter information is fortunately available as each selected household is assigned a unique Sample Unit ID (SUID). Although the correlation between members of the same household is not explicitly modeled in (3.2), it is accounted for in the proposed variance estimators of chapter 3.

Interviews were conducted at four-month intervals; this resulted in a recall bias commonly referred to as the seam effect, which was discussed in section 2.1. The most common methods to adjust for the seam effect introduced by an individual that reported a transition at the seam between two waves of interviews involve reallocating that individual's sampling weight to all of the months between the two interviews in question; see Kalton et al. (1992), section 2.2. As the methods proposed in chapter 3 are meant to be unweighted and seam effect corrections are based on the reallocation of sampling weights, no seam effect adjustment was made in our analyses. Ignoring the seam effect should have little impact on inference for β_0 in (3.2), but estimation of $\Lambda_{0h}(t)$ and $\hat{S}_h(t)$ is affected. As can be seen in figure 2.1, the seam effect persists even though spell durations were converted from a monthly to a daily basis, as described in the first paragraph of section 4.1.

There exist various definitions of what a spell is in SIPP. For example, is one month off the food stamp program considered as the termination of a spell, or should a longer time interval be allowed to account for administrative or other technical delays. Evidence of gaps in program participation is given by Kalton et al. (1992), section 3.4. They noted that, with one month off the program defining the end, 11% of individuals completing a spell on the food stamp program return to the program within two months, 19% return within four months and 25% within a year. Some of the effects of allowing different time gaps (i.e., none, 1, 2, 4 and 6 months) in a spell are shown in tables 3.4 and 3.7 of Kalton et al. (1992). Unfortunately, they do not give any strong argument in favor of any given definition. In agreement with the final analyses done by Kalton et al. (1992) in their chapter 4, no gap was allowed in defining spell duration. This definition was applied to both the analyses of spells on the food stamps program in section 4.1.1 and of spells without health insurance in section 4.1.2.

4.1.1 Food stamps

Following the discussion of section 3.3.2, three analyses based on model (3.2) were performed in section 4.1.1. The first two are weighted and unweighted analyses based on the individuals with strictly positive sampling weights only, and the last one is an unweighted analysis based on all individuals (i.e., both individuals with strictly positive sampling weights and individuals with zero as sampling weights). The results are shown in table 4.1, with standard deviations given in parentheses underneath the point estimates.

Table 4.1, column 1 contains the results of a weighted stratified marginal analy-

$\hat{\boldsymbol{\beta}}$	**weighted**	**unweighted**	
	$n = 1,095$	$n = 1,095$	$n = 1,816$
$I(\text{Sex})$	-1.41×10^{-1}	-1.76×10^{-1}	-1.60×10^{-1}
	$(6.03 \times 10^{-2})\dagger$	$(5.73 \times 10^{-2})\ddagger$	$(4.59 \times 10^{-2})\ddagger$
$I(\text{Indian})$	5.83×10^{-1}	6.30×10^{-1}	4.71×10^{-1}
	(2.87×10^{-1})	(2.73×10^{-1})	(2.18×10^{-1})
Age	-6.33×10^{-3}	-7.01×10^{-3}	-4.62×10^{-3}
	(2.44×10^{-3})	(2.27×10^{-3})	(1.83×10^{-3})
Education	-2.60×10^{-2}	-2.77×10^{-2}	-1.06×10^{-2}
	(1.46×10^{-2})	(1.43×10^{-2})	(1.17×10^{-2})
Income	-3.17×10^{-4}	-3.54×10^{-4}	-2.59×10^{-4}
	(2.95×10^{-4})	(2.86×10^{-4})	(2.11×10^{-4})
$\text{Income} \times I(\text{Income} > 500)$	6.59×10^{-4}	6.60×10^{-4}	4.81×10^{-4}
	(2.28×10^{-4})	(2.25×10^{-4})	(1.80×10^{-4})
$\text{Education} \times I(\text{Education} > 10)$	2.12×10^{-2}	2.25×10^{-2}	1.15×10^{-2}
	(9.32×10^{-3})	(9.28×10^{-3})	(7.34×10^{-3})

† s.d.'s computed using the method of Lin (2000), section 3

‡ s.d.'s computed using $\widehat{V}_R(\hat{\beta})$, given by (3.19)

Table 4.1: Weighted and unweighted marginal Cox PH analyses for spells on the food stamps program.

sis based on 1,095 uncensored and right-censored spells on the food stamps program experienced by individuals with strictly positive weights only. This analysis uses the design-based methodology of Lin (2000) for super-population inference. The analysis in column 2 is the unweighted version of the one in column 1, and was done using the methodology proposed in chapter 3. In column 3, the sample of the unweighted analysis of column 2 is expanded to include all individuals; it is based on 1,816 uncensored and right-censored spells on the food stamps program.

Looking at table 4.1, point estimates for β_0 for the three columns are similar. However, all of the positive estimates of column 3 are smaller that the ones of the other two columns and, except for $I(\text{Sex})$, the negative estimates of column 3 are larger that the ones of the other two columns. This might indicate that individuals that joined existing households after the start of the panel in 1987 behaved slightly differently. Nevertheless, all of these point estimates are not statistically different at the 10% level when carrying pairwise comparisons. This statement holds regardless of the variance estimator chosen (e.g., Lin's (2000) or $\widehat{V}_R(\hat{\beta})$ when comparing columns 1 and 2) to do the 21 different pairwise comparisons. Unfortunately, when making these pairwise comparisons between the point estimates of columns 1 and 2, the variance terms that should be used as denominators are not Lin's (2000) variance estimator or $\widehat{V}_R(\hat{\beta})$ alone, but a combination of the two and of their covariance. As the main goal of the examples given in chapter 4 is to illustrate the methods proposed in chapter 3 and that developing an exact test for these pairwise comparisons was outside the scope of this thesis, this simple ad hoc procedure was preferred. Similar comments also apply to comparisons between point estimates of columns 2 and 3, and to the test given by (3.68).

Similarities between columns 1 and 2 imply that the assumption of uninformative and/or ignorable sampling designs are satisfied. In addition, this also justifies the choice and use of model (3.2), with the covariates and strata described in section 4.1 and in table 4.1. Similarities between columns 2 and 3 imply that individuals that joined existing households after the start of the study in 1987 (i.e., individuals with zero sampling weights) are not different from individuals that were part of the original sample, with respect to the duration of spells on the food stamps program and its relation to the covariates of table 4.1.

Standard deviations in column 2 are slightly smaller than the ones in column 1. This is due to the fact that unweighted procedures are more efficient, when the sampling design is uninformative; see section 3.3.2. However, the real gain is that the methods proposed in chapter 3 naturally extend to include all individuals regardless of their sampling weights. This can be seen by looking at the standard deviations shown in column 3.

The only race statistically different from the white/Caucasian baseline is American Indian, Eskimo or Aleut (i.e., I(Indian)). Hence, the other race covariates (i.e., I(Black) and I(Asian)) are not included in the final model. Except for education and income, all the variables of table 4.1 are statistically significant at the 5% level for all three models. The reasons for including education and income regardless of that are that both Education$\times I$(Education > 10) and Income$\times I$(Income > 500) are statistically significant. In particular, the latter variable is significant at the 1% level for all three models. The purpose of Education$\times I$(Education > 10) is to allow for a different relationship between duration of spells on the food stamp program and years of education for individuals with more that 10 years of education. Similarly, Income$\times I$(Income > 500) allows for a different relation for individuals earning more than \$500 a month. Part of the motivation behind including Education$\times I$(Education > 10) in the model is given in figures 4.1 and 4.2 (figure 4.2 being a zoom on the smooth line of figure 4.1). Both figures examine the functional form of education in the model of table 4.1, column 2, without the last two covariates by plotting the martingale residuals, as defined by (4.2), versus years of education. A smooth regression line, obtained from S-Plus loess.smooth function, is also drawn to help in identifying a possible functional form.

The martingale residual process corresponding to the i^{th} individual of the h^{th} stratum is

$$\hat{M}_{hi}(t) = N_{hi}(t) - \int_0^t \gamma_{hi}(u)\ \exp\{\hat{\beta}' x_{hi}(u)\}\, d\hat{\Lambda}_{0h}(u, \hat{\beta})\ , \tag{4.1}$$

where $\gamma_{hi}(t)$ and $\hat{\Lambda}_{0h}(t, \hat{\beta})$ were, respectively, defined in (3.3) and (3.34). Recall that the vector $\hat{\beta}$ is the maximum partial likelihood estimate defined as the solution of $U(\beta) = 0$, where $U(\beta)$ was given in (3.7). As there is no time-varying covariates in any of the Cox models fitted in chapter 4, (4.1) simplifies and is called the martingale residual, given by

$$\hat{M}_{hi}(t_{hi}) = \delta_{hi} - \exp\{\hat{\beta}' x_{hi}(t_{hi})\}\ \hat{\Lambda}_{0h}(t_{hi}, \hat{\beta})\ . \tag{4.2}$$

See Therneau & Grambsch (2000), section 4.2, and Klein & Moeschberger (1997), section 11.3, for further information on martingale residuals and their use to determine the functional form of a covariate.

Going back to figures 4.1 and 4.2, the critical points at 10 and 16 years of education on figure 4.2 are an indication of changes in the relationship between duration of spells on the food stamp program and years of education. Looking back at the data itself or at figure 4.1, few individuals have more than 16 years of education and including an extra variable in the model to account for these individuals did not yield anything that was statistically significant. Note that one could argue in favor of including a categorical variable, with the following three levels: 1 — less than high school degree, 2 — completed high school, and 3 — completed university/college degree, instead of years of education. As our goal was to illustrate the methods of chapter 3, using years of education, a continuous variable was preferred. Similarly, justification for including Income$\times I(\text{Income} > 500)$ in the model can be found by looking at figures 4.3 and 4.4. Figure 4.3 is a plot of martingale residuals, as defined by (4.2), versus monthly income and figure 4.4 is a zoom on the smooth line of figure 4.3. As food stamp is a program geared towards helping poor people the change in the relationship between spell duration and income and the shape of the curve in figure 4.4 are not surprising.

The following conclusions on the effects of the various covariates can be reached from table 4.1:

- Women are more likely to experience longer spells on the food stamp program;
- Indians are more likely to experience shorter spells on the food stamp program (this might be because they are eligible for other forms of help);
- Older individuals are more likely to experience longer spells on the food stamp program;
- For individuals with 10 years of education or less, education has no effect on spell duration;

- For individuals with more than 10 years of education, more education implies that they are more likely to experience shorter spells on the food stamp program (although, the effect is marginal);

- For individuals earning \$500 per month or less, income has no effect on spell duration;

- For individuals earning more than \$500 per month, a higher income implies that they are more likely to experience shorter spells on the food stamp program (individuals making more than the maximum income allowed to receive food stamps stop to be eligible).

Various standard deviation estimates, including the ones of table 4.1, are presented in table 4.2. The five columns of table 4.2 contain the following standard deviation estimates: column 1 — Lin (2000) design-based estimates for super-population inference; columns 2 and 3 — proposed estimates of chapter 3 given by (3.19) and (3.75); and columns 4 and 5 — estimates which assume that all the observations are independent, also referred to as "naive" s.d. estimates. Column 2 is based on individuals with strictly positive sampling weights (i.e., does not include top-ups), while column 3 is based on all individuals who experienced spells on the food stamp program (i.e., includes top-ups); similarly for column 4 versus 5. Standard deviation estimates from Binder's (1992) method and from (3.75), the weighted version of (3.19), are not shown as they are practically identical to the ones of column 1, to the accuracy shown in the table.

For all the tables of chapter 4 where they are present, Binder's (1992) and Lin's (2000) variance estimates were computed using SUDAAN SURVIVAL procedure[2]. Note that the strhaz statement is required to fit stratified Cox models as given by (3.2). The

[2] To compute Lin's (2000) variance estimate for super-population inference, as described in section 3 of his article, an extra term must be added to Binder's (1992) variance estimate. This term is just the "naive" variance estimate obtained by fitting a weighted stratified Cox model; it is easily obtained using SAS or S-Plus.

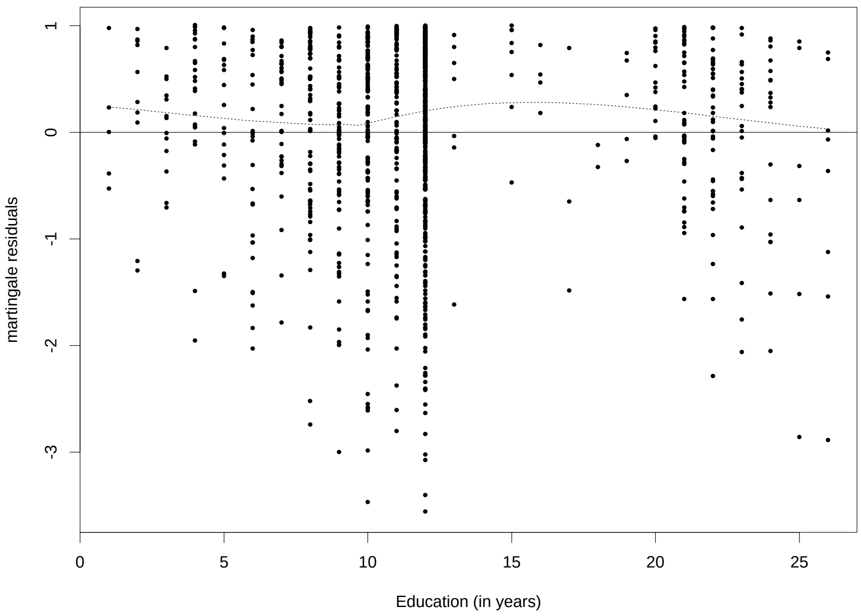

Figure 4.1: Martingale residuals vs. Education — model of table 4.1, column 2, without the last two covariates.

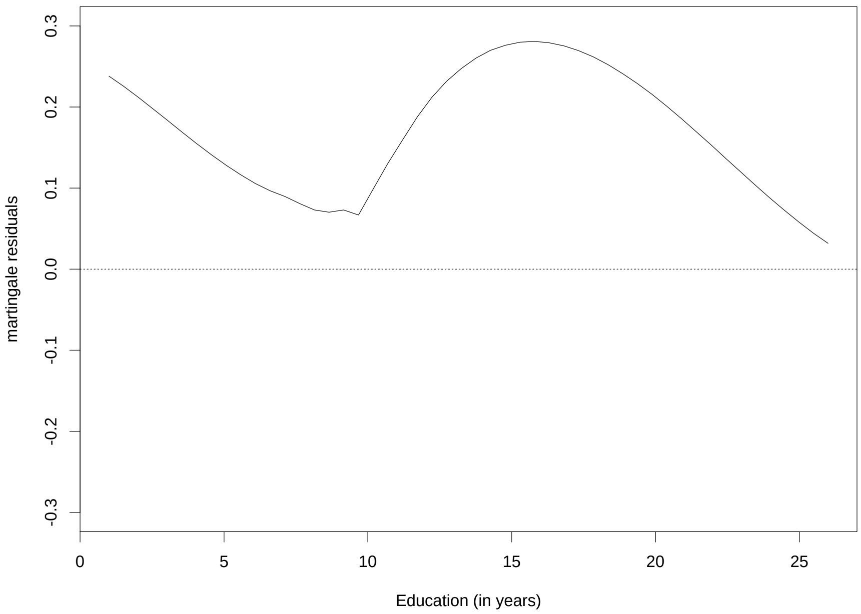

Figure 4.2: Martingale residuals — zoom of figure 4.1.

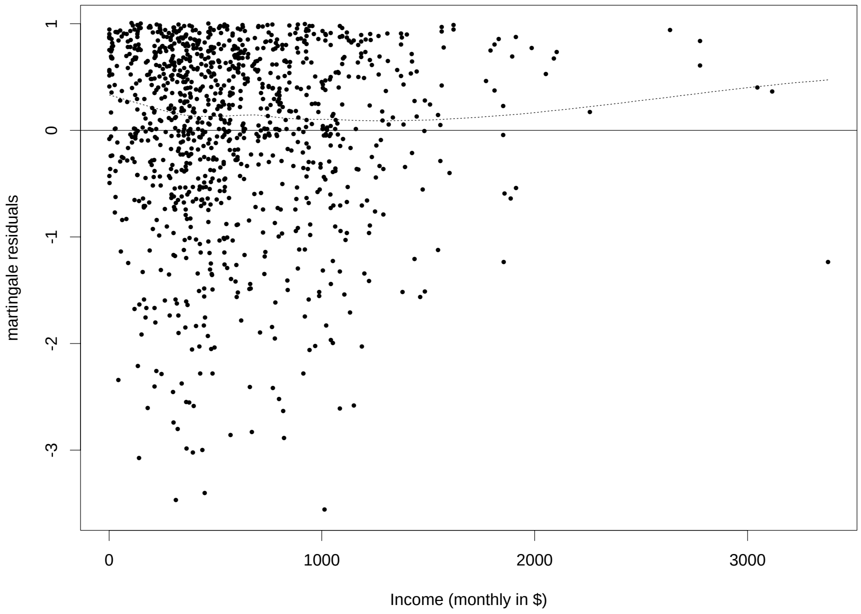

Figure 4.3: Martingale residuals vs. Income — model of table 4.1, column 2, without the last two covariates.

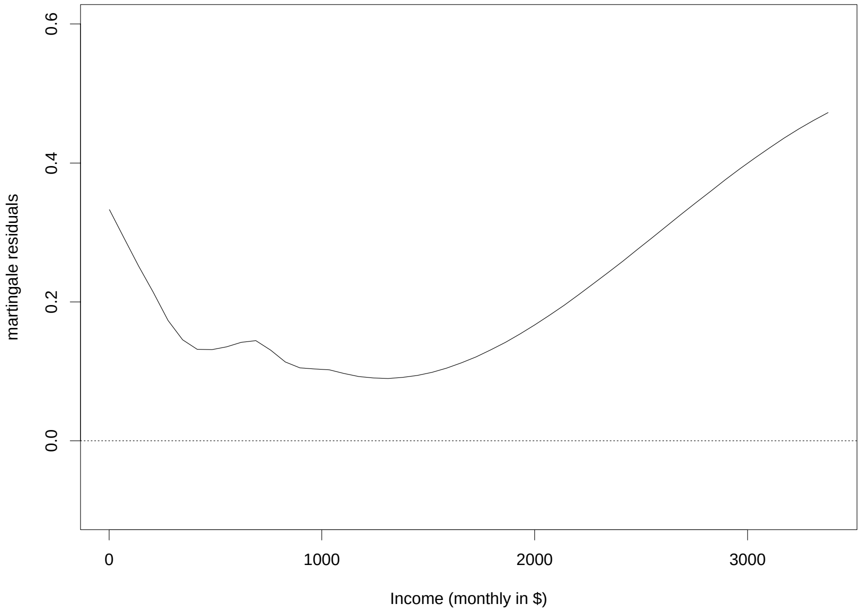

Figure 4.4: Martingale residuals — zoom of figure 4.3.

$\hat{\boldsymbol{\beta}}$	**weighted**	**unweighted**			
	$\sqrt{\widehat{V}_{LIN}(\hat{\beta})}$	$\sqrt{\widehat{V}_R(\hat{\beta})}$		$\sqrt{\mathcal{I}(\hat{\beta})^{-1}}$	
	$n = 1,095$	$n = 1,095$	$n = 1,816$	$n = 1,095$	$n = 1,816$
I(Sex)	6.03×10^{-2}	5.73×10^{-2}	4.59×10^{-2}	7.36×10^{-2}	5.69×10^{-2}
I(Indian)	2.87×10^{-1}	2.73×10^{-1}	2.18×10^{-1}	2.58×10^{-1}	2.12×10^{-1}
Age	2.44×10^{-3}	2.27×10^{-3}	1.83×10^{-3}	2.34×10^{-3}	1.84×10^{-3}
Education	1.46×10^{-2}	1.43×10^{-2}	1.17×10^{-2}	1.59×10^{-2}	1.25×10^{-2}
Income	2.95×10^{-4}	2.86×10^{-4}	2.11×10^{-4}	3.02×10^{-4}	2.13×10^{-4}
Income$\times I$(Income > 500)	2.28×10^{-4}	2.25×10^{-4}	1.80×10^{-4}	2.39×10^{-4}	1.81×10^{-4}
Education$\times I$(Education > 10)	9.32×10^{-3}	9.28×10^{-3}	7.34×10^{-3}	1.02×10^{-2}	7.77×10^{-3}

Table 4.2: Standard deviation estimates for the different weighted and unweighted marginal Cox PH analyses for spells on the food stamps program.

keywords beta and covar of the print statement are also practically essential as SUDAAN, by default, gives the results for both $\hat{\beta}$ and $\widehat{V}(\hat{\beta})$ with only two decimal point precision. In addition, the ties=breslow option of the model statement should be used instead of the default one, ties=efron, which has known computer bugs. See SUDAAN User's Manual (Research Triangle Institute (2001)) for further information on the SURVIVAL procedure and the various methods to enter data into SUDAAN.

As mentioned before, the estimates of column 2 are slightly smaller than the ones of column 1 as the unweighted analysis is more efficient when the sampling design is uninformative and/or ignorable. Column 3 estimates are the smallest as they are both unweighted and based on more spells. Columns 4 and 5 estimates, which are based on the incorrect independence assumption, are generally larger than the corresponding ones which account for correlation between individuals and spells. However, they are not extremely different from those corresponding estimates, indicating a small effect due to association of spells on the food stamp program for individuals living in the same household and for individuals experiencing multiple spells. Part of this is explained by the fact that numerous children were not included in the analyses of section 4.1.1 and that 79.8% of all benefits went to households with children. Note that 39.6% of them were headed by a single parent, the overwhelming majority of whom were women. For more information on characteristics of food stamp households, see the web site of the Food and Nutrition Service (FNS), U.S. Department of Agriculture, at `www.fns.usda.gov/fsp/`.

Diagnostic model checking is straightforward for the unweighted analyses; see Therneau & Grambsch (2000), chapters 4 to 7. However, research is needed to account for the effects of correlation between individuals and spells on diagnostic plots and on other statistics. In this case fortunately, the effect is probably marginal as the "naive" standard deviation estimates of table 4.2 are close to the ones that account for correlation; as was discussed in the previous two paragraphs. Figures 4.5 and 4.6, as well as other graphics and diagnostic procedures, did not reveal any serious departure from model (3.2), with the strata and covariates described in sections 4.1 and 4.1.1. Figures 4.5 and 4.6 were obtained from S-Plus cox.zph function, which tests the proportional hazards assumption using the method introduced by Grambsch & Therneau (1994). Strangely, no individual from New Hampshire experienced a spell on the food stamp program during the 28 months followup period. Unfortunately, we can not provide any good explanation for this phenomenon. Perhaps New Hampshire has a separate program.

4.1.2 Health insurance

As in section 4.1.1, three analyses based on model (3.2) were performed in this section. The first two are weighted and unweighted analyses based on individuals with strictly positive sampling weights only and the last one is an unweighted analysis based on all individuals. The results are shown in table 4.3, with standard deviations given in parentheses.

Table 4.3 has the same features as table 4.1. Hence, column 1 contains the results of a weighted stratified marginal analysis based on 4,314 uncensored and right-censored spells without health insurance experienced by individuals with strictly positive weights only. This analysis uses the design-based methodology of Lin (2000) for super-population inference. The analysis in column 2 is the unweighted version of the one in column 1, and was done using the methodology proposed in chapter 3. In column 3, the sample of the analysis of column 2 is expanded to include all individuals; it is based on 7,656 uncensored and right-censored spells. As in table 4.1, point estimates for β_0 for the three columns of table 4.3 are not statistically different at the 10% level (except age which is at the 5% level). This statement holds regardless of the variance estimator chosen to do the different pairwise comparisons. When using the Bonferroni correction for multiple comparisons, age is also not statistically different at the 10% level. Comments made earlier (see discussion following table 4.1) regarding the choice of the denominators used in these pairwise comparisons also apply here.

The strong similarities between columns 1 and 2 imply that the assumption of uninformative and/or ignorable sampling designs are satisfied. In addition, this also justifies

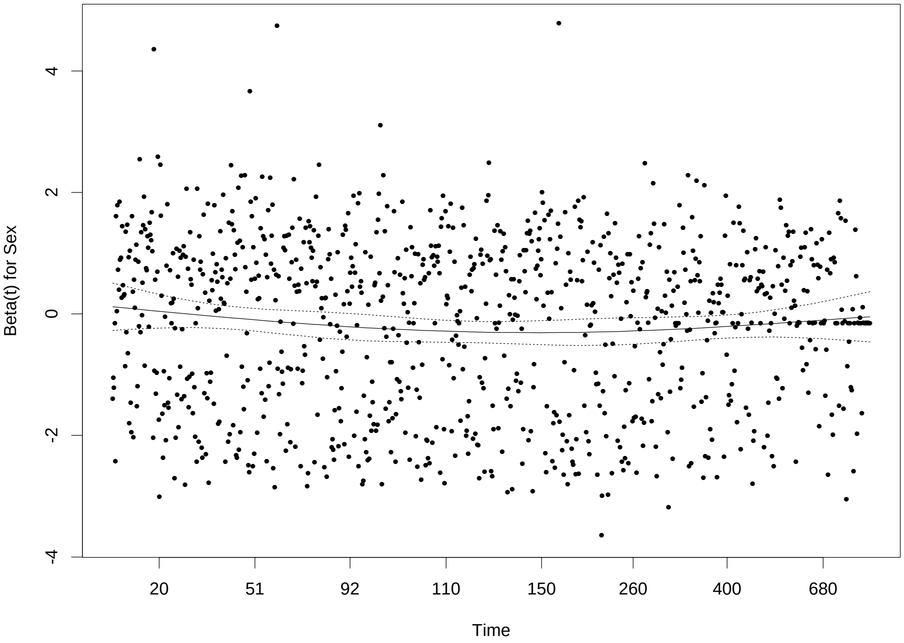

Figure 4.5: Test of PH assumption for I(Sex) in model of table 4.1, column 2.

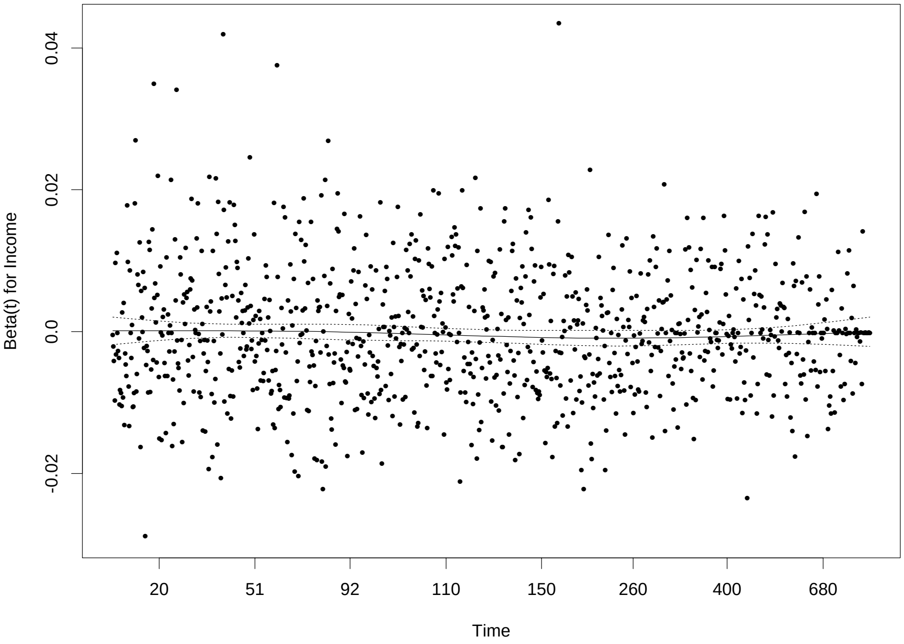

Figure 4.6: Test of PH assumption for Income in model of table 4.1, column 2.

$\hat{\boldsymbol{\beta}}$	**weighted**	**unweighted**	
	$n = 4,314$	$n = 4,314$	$n = 7,656$
I(Sex)	7.04×10^{-2}	7.45×10^{-2}	8.92×10^{-2}
	$(3.31 \times 10^{-2})\dagger$	$(3.19 \times 10^{-2})\ddagger$	$(2.46 \times 10^{-2})\ddagger$
Age	7.44×10^{-3}	6.10×10^{-3}	4.33×10^{-3}
	(2.19×10^{-3})	(2.17×10^{-3})	(1.75×10^{-3})
Education	2.39×10^{-2}	2.02×10^{-2}	1.66×10^{-2}
	(7.45×10^{-3})	(7.46×10^{-3})	(5.81×10^{-3})
Income	-3.42×10^{-4}	-3.39×10^{-4}	-3.27×10^{-4}
	(1.07×10^{-4})	(1.03×10^{-4})	(7.94×10^{-5})
Age × Education	-3.89×10^{-4}	-3.23×10^{-4}	-2.49×10^{-4}
	(1.65×10^{-4})	(1.64×10^{-4})	(1.35×10^{-4})
Income$\times I$(Income > 600)	4.45×10^{-4}	4.35×10^{-4}	4.27×10^{-4}
	(9.77×10^{-5})	(9.43×10^{-5})	(7.23×10^{-5})

† s.d.'s computed using the method of Lin (2000), section 3

‡ s.d.'s computed using $\widehat{V}_R(\hat{\beta})$, given by (3.19)

Table 4.3: Weighted and unweighted marginal Cox PH analyses for spells without health insurance.

the choice and use of model (3.2), with the covariates and strata described in section 4.1 and in table 4.3. Similarities between columns 2 and 3 imply that individuals that joined existing households after the start of the study in 1987 are not different from individuals that were part of the original sample, with respect to the duration of spells without health insurance and its relation to the covariates of table 4.3.

Standard deviations in column 2 are slightly smaller than the ones in column 1, except for education. Again, this is due to the fact that unweighted procedures are more efficient when the sampling design is uninformative. However, the real gain is the ability to include all individuals, when using the proposed unweighted methods of chapter 3. This can be seen by looking at the standard deviations shown in column 3.

None of the race variables are statistically different from the white/Caucasian baseline, and are not included in the final model. Except for the Age×Education interaction in the model of column 3 (which has a p-value = 6.5%), all the covariates of table 4.3 are statistically significant at the 5% level for all three models. With a p-value $< 0.1\%$ for all three models, Income$\times I$(Income > 600) is highly statistically significant. However, the corresponding $\hat{\beta}$'s are small and the effect is marginal. The purpose of Income$\times I$(Income > 600) is to allow for a different relationship between duration of spells without health insurance and monthly income for individuals earning more than \$600 a month. The motivation for including Income$\times I$(Income > 600) in the model is given in figures 4.7 and 4.8 (figure 4.8 being a zoom on the smooth line of figure 4.7). As figures 4.1 and 4.3, figure 4.7 is a plot of martingale residuals, as defined by (4.2), versus the covariate of interest, monthly income in the current case. The critical point at about \$600 on figure 4.8 indicates a change in the relationship between duration of spells without health insurance and monthly income. This is not surprising as poor people are eligible to various social programs including Medicaid, a jointly funded Federal–State health insurance program for certain low-income and needy people.

The following conclusions on the effects of the various covariates can be reached from table 4.3:

- Women are more likely to experience shorter spells without health insurance;
- Older individuals are more likely to experience shorter spells without health insurance (many individuals aged 65 and older are eligible for Medicare health coverage);
- Individuals with more education are more likely to experience shorter spell without health insurance;
- For individuals earning \$600 per month or less, less money implies that they are

more likely to experience shorter spells without health insurance (low income individuals are eligible for Medicaid health coverage);

- For individuals earning more than \$600 per month, a higher income implies that they are more likely to experience shorter spells without health insurance (although, the effect is marginal).

Various standard deviation estimates, including the ones of table 4.3, are presented in table 4.4. The five columns of table 4.4 contain the same estimators as the ones of table 4.2; that is: column 1 — Lin (2000) design-based estimates for super-population inference; columns 2 and 3 — proposed estimates given by (3.19) and (3.75); and columns 4 and 5 — estimates which assumes that all the observations are independent, also referred to as "naive" s.d. estimates. Columns 2 and 4 are based on individuals with strictly positive sampling weights only, while columns 3 and 5 include all individuals who experienced spells without health insurance. Standard deviation estimates from Binder's (1992) method and from (3.75) are not shown as they are practically identical to the ones of column 1, to the accuracy shown in the table.

As mentioned before, the estimates of column 2 are slightly smaller than the ones of column 1 as the unweighted analysis is more efficient when the sampling design is uninformative and/or ignorable. Column 3 estimates are the smallest as they are both unweighted and based on more spells. Columns 4 and 5 estimates, which are based on the incorrect independence assumption, are not extremely different from the ones which account for correlation between individuals and spells. This is an indication of a small effect due to association of spells without health insurance.

For the unweighted analyses, the same diagnostic model checking tools as the ones used at the end of section 4.1.1 can also be performed here. Again, research is needed to account for the effects of correlation between individuals and spells on diagnostic plots and other statistics. Fortunately, the effect is probably marginal in this case as the "naive" standard deviations estimates of table 4.4 are close to the ones that account for

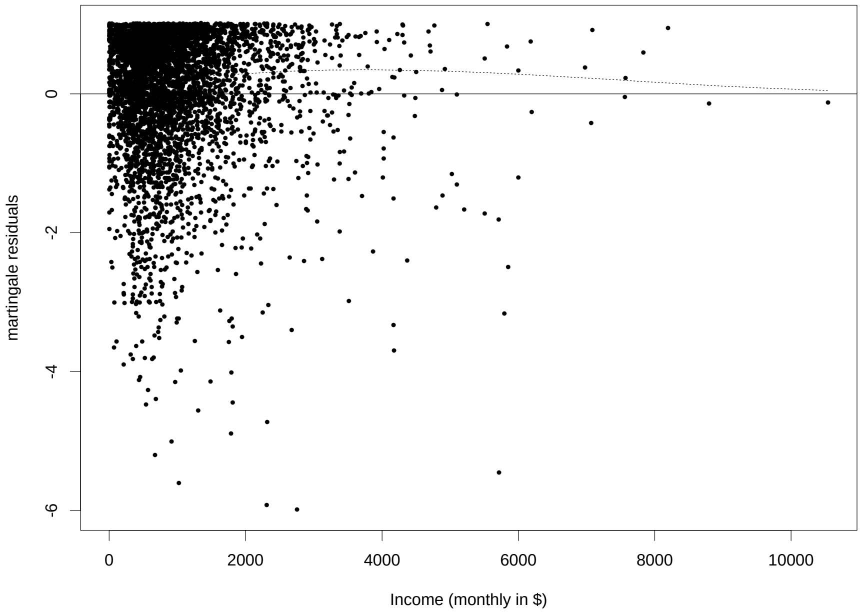

Figure 4.7: Martingale residuals vs. Income — model of table 4.3, column 2, without the last two covariates.

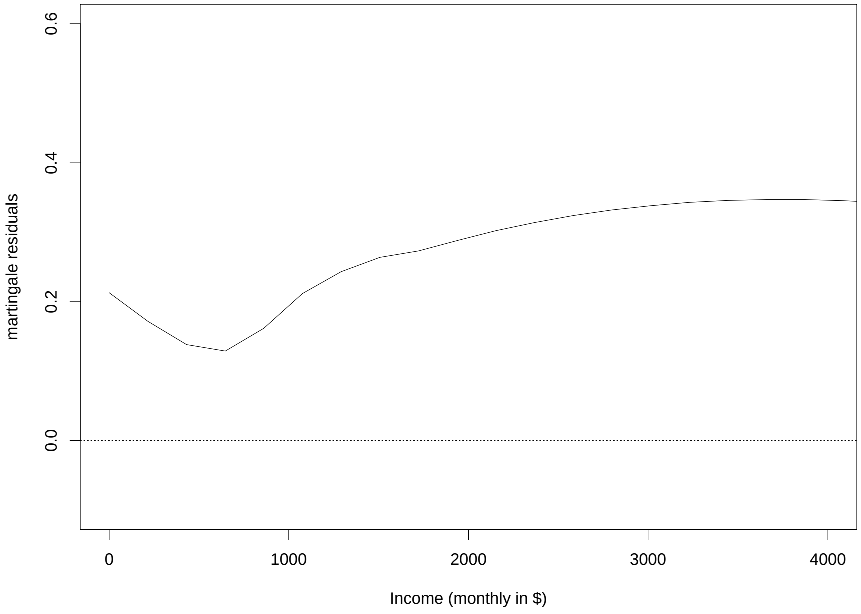

Figure 4.8: Martingale residuals — zoom of figure 4.7.

$\hat{\boldsymbol{\beta}}$	**weighted**	**unweighted**			
	$\sqrt{\widehat{V}_{LIN}(\hat{\beta})}$	$\sqrt{\widehat{V}_{R}(\hat{\beta})}$		$\sqrt{\mathcal{I}(\hat{\beta})^{-1}}$	
	$n = 4,314$	$n = 4,314$	$n = 7,656$	$n = 4,314$	$n = 7,656$
I(Sex)	3.31×10^{-2}	3.19×10^{-2}	2.46×10^{-2}	3.47×10^{-2}	2.65×10^{-2}
Age	2.19×10^{-3}	2.17×10^{-3}	1.75×10^{-3}	2.06×10^{-3}	1.65×10^{-3}
Education	7.45×10^{-3}	7.46×10^{-3}	5.81×10^{-3}	6.82×10^{-3}	5.33×10^{-3}
income	1.07×10^{-4}	1.03×10^{-4}	7.94×10^{-5}	1.05×10^{-4}	8.05×10^{-5}
Age × Education	1.65×10^{-4}	1.64×10^{-4}	1.35×10^{-4}	1.52×10^{-4}	1.24×10^{-4}
Income×I(Income > 600)	9.77×10^{-5}	9.43×10^{-5}	7.23×10^{-5}	9.62×10^{-5}	7.35×10^{-5}

Table 4.4: Standard deviation estimates for the different weighted and unweighted marginal Cox PH analyses for spells without health insurance.

correlation. Figures 4.9 and 4.10, as well as other graphics and diagnostic procedures, did not reveal any serious departure from the model.

4.2 SLID analyses

The data for the 1st SLID panel, which ran from 1993 to 1998, was obtained through Statistics Canada's Research Data Centres (RDC) program. The analyses of section 4.2 were carried out at the South-Western Ontario Research Data Centre (SWORDC), located at the University of Waterloo. The RDC program is part of Statistics Canada's Data Liberation Initiative (DLI), which aimed to provide Canadian academic institutions with affordable access to Statistics Canada data files and databases for teaching and research purposes. The RDC is an initiative of Statistics Canada, the Social Sciences and Humanities Research Council (SSHRC) and the consortium of Canadian universities. For more information on the RDC program, see `www.statcan.ca/english/rdc/`. The 1993–1998 SLID dataset is even larger than the SIPP dataset of section 4.1 and is split in yearly sub-datasets. To extract the relevant variables and to match the appropriate records, both SLIDRET and SAS were used. The former is a software designed by Statistics Canada and is based on Visual FoxPro. The name SLIDRET stands for SLID Data Retrieval System and the SLIDRET User's Manual can be freely downloaded from the internet.

In contrast to SIPP where spell duration is not uniquely defined, jobless spells are well defined in SLID. As mentioned in section 1.1.1, a jobless spell refers to a period in which an individual had no attachment of any type to an employer and had worked previously. It is considered to have ended when any type of employment is identified after the start of the layoff period. SLIDRET gives the duration in weeks of the jobless spells. Hence, any spell that lasted less than a week is not included in the dataset. One of the reasons to convert from a daily basis (which is available in SLID as the start and end dates of every jobless spell are recorded) to a weekly basis is to minimize recall errors.

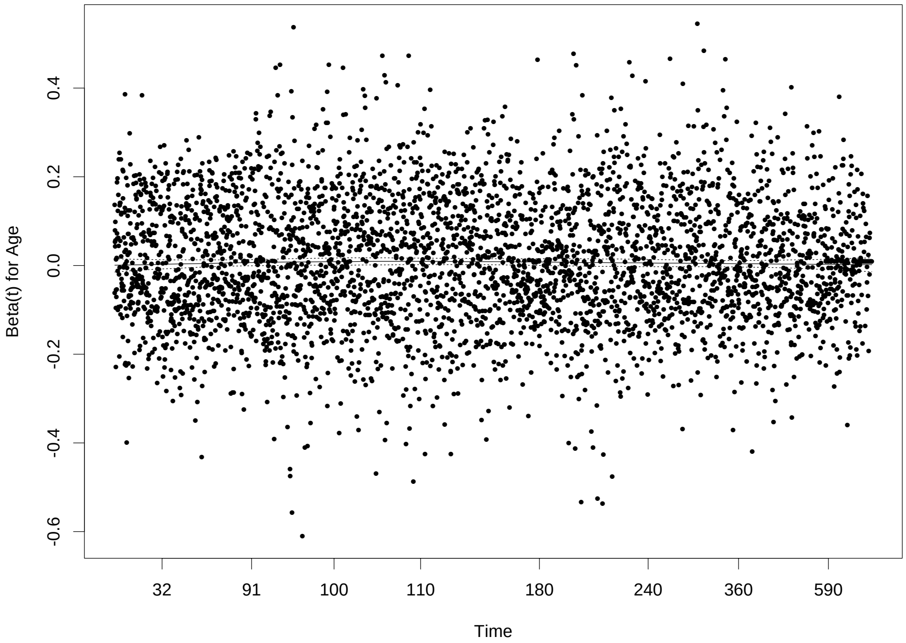

Figure 4.9: Test of PH assumption for Age in model of table 4.3, column 2.

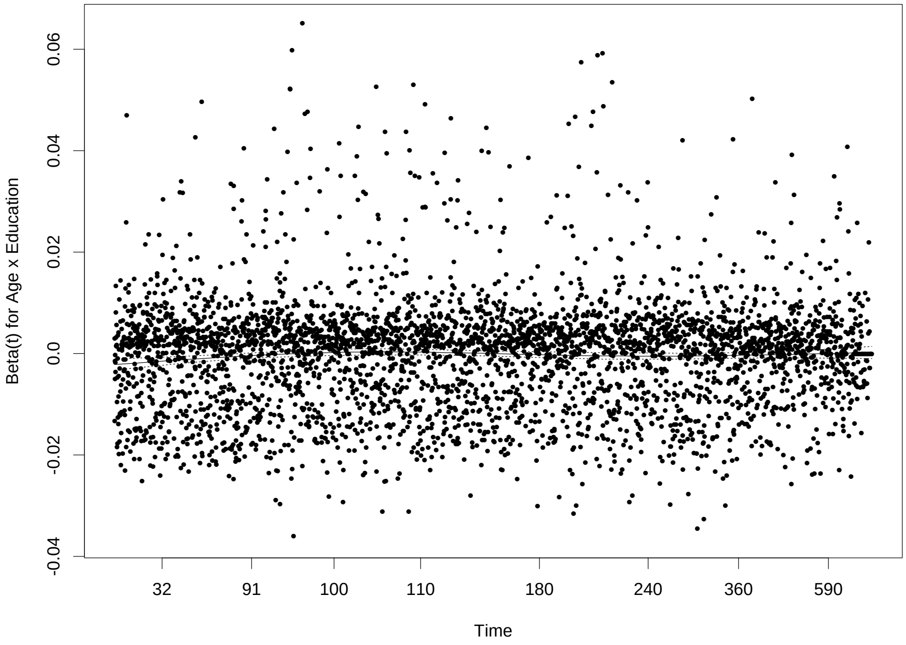

Figure 4.10: Test of PH assumption for Age × Education in model of table 4.3, column 2.

As in sections 4.1.1 and 4.1.2, stratified Cox PH models, as given by (3.2), were fitted to the duration (in weeks) of jobless spells in section 4.2.1. These models are composed of 11 strata, each having its own baseline hazard function $\lambda_{0h}(t)$, $h = 1, \ldots, 11$. These strata correspond to the 10 Canadian provinces and another stratum containing individuals who could not be classified in any of the 10 provinces regardless of the reasons. This last stratum includes individuals living in the 3 Canadian territories, the United-States and other countries.

The following covariates were included in the exponential term of model (3.2):

- Sex: man (baseline) and woman; that is,

$$I(\text{Sex}) = \begin{cases} 1 & \text{if the individual is a woman} \\ 0 & \text{otherwise ;} \end{cases}$$

- Looked for job during jobless spell: looked and did not looked (baseline); that is,

$$I(\text{Looked}) = \begin{cases} 1 & \text{if the individual looked for a job during jobless spell} \\ 0 & \text{otherwise ;} \end{cases}$$

- Children: has one or more children and has none (baseline); that is,

$$I(\text{Children}) = \begin{cases} 1 & \text{if the individual has one or more children} \\ 0 & \text{otherwise ;} \end{cases}$$

- Age, which is defined as age in years (as of January 1, 1993) $-$ 28.58 (the mean);
- Education or years of school attended;
- Hourly-wage (in dollars) earned by the individual in the job hold prior to the jobless spell;
- Employment insurance (EI): received EI and did not (baseline); that is,

$$I(\text{EI}) = \begin{cases} 1 & \text{if the individual received EI during jobless spell} \\ 0 & \text{otherwise ;} \end{cases}$$

- Visible minority: is a member of a visible minority and is not (baseline); that is,

$$I(\text{Minority}) = \begin{cases} 1 & \text{if the individual reported being a member} \\ & \text{of a visible minority} \\ 0 & \text{otherwise ;} \end{cases}$$

- Winter: jobless spell began in the winter (i.e., from November 1 to March 31) and did not began in the winter (baseline); that is,

$$I(\text{Winter}) = \begin{cases} 1 & \text{if the individual began his/her jobless spell in the winter} \\ 0 & \text{otherwise .} \end{cases}$$

The reason for defining the covariate age as above, is to reduce collinearity when fitting the three models of table 4.6, which include an age square term. Defining winter as the period from November 1 to March 31 is somewhat arbitrary. This was done to account for seasonal workers, which do not return to work for the same employer; see theme 9 on page 6. In addition, the interaction term $I(\text{Winter}) \times I(\text{Weeks} > 52)$ is an indicator, which takes value 1 if the individual began his/her jobless spell in the winter and if that spell lasted more than 52 weeks and takes value 0 otherwise. As can be seen in tables 4.6 and 4.8, both $I(\text{Winter})$ and $I(\text{Winter}) \times I(\text{Weeks} > 52)$ are highly significant in all the Cox models we fitted to the SLID jobless dataset. Moreover, other authors also choose to include a similar winter covariate.

The other three variables that are part of the analyses of section 4.2.1 are the longitudinal sampling weights and two different variables used to define clusters: the household ID and the enumeration area (EA) number. In addition to the longitudinal sampling weights computed at the start of the 1^{st} panel, new weights are computed at the end of every year for the duration of the panel. These weights are adjustments of the original ones that account for attrition and other types of non-response. As discussed in sections 2.6 and 3.3.2, defining and computing longitudinal sampling weights is a complex process. Moreover, which of the six yearly sampling weights should be used? For example, if a jobless spell, that started in 1993, ended in 1995 should the longitudinal sampling weight computed at the end of 1995 be used or the one computed at the end of the study in 1998. A more complex situation arises when a jobless spell started in 1994 and ended in 1995. In this case, should the longitudinal sampling weight computed at the end of 1994, 1995 or 1998 be used in weighted analyses? Another possibility is to use the original longitudinal sampling weights assigned at the beginning of the study in 1993. Using either the sampling weight of the year in which the spell started or ended implies that longitudinal sampling weights from different years would be combined together in the analysis. Hence, they would no longer represent a given population at a given time, but rather an awkward mix of that population over the years it was under study. The original longitudinal sampling weights are a better choice. Unfortunately, it is likely that they no longer represent the population after years of attrition and other forms of non-response. As this is not the case for the longitudinal sampling weight computed at the end of the study, they are probably the best choice. This also corresponds to the sampling weights given by SLIDRET by default. However, this is not a perfect solution and more research is needed on the subject; see section 6.2. Among the total of 16,682 jobless spells experienced by individuals from the 1st panel, 14,978 or about 90% of them were experienced by individuals that were part of the original sample selected in January 1993 and that were still under study at the end of 1998. These individuals are the only ones with strictly positive longitudinal sampling weights; the others have a sampling weight of zero.

As explained at the beginning of section 1.1.1, there are many levels of stratification and clustering in SLID. In addition, some individuals experienced multiple spells of unemployment, which could also be considered as an extra level of clustering. Household ID is a unique identifier for the household to which the respondent belonged as of December 31 of the reference year. Using it to define clusters will account for correlation between individuals living together and, to a certain degree, for the correlation resulting from individuals experiencing multiple jobless spells. Every individual can be a member of only one household in a given year. However, individuals can move to different households each year. Hence, after the six year follow-up period, some individuals will have lived in multiple households. Since no statistical method can handle multiple membership to clusters at the same level of clustering, a unique household must be defined for every individual. For the analyses of section 4.2.1, this was done by assigning the household in which the individual, who lived in multiple households, spent the greatest amount of time over the life of the panel. This resulted in the 16,682 jobless spells being divided into 9,641 different household ID's.

Enumeration areas (EA's) are geographic areas, defined by the 1991 Census Geography, covered by individual enumerators for the census. EA codes are not unique across provinces and federal electoral districts. To obtain a unique value the fields corresponding to the province of residence, the federal electoral district and the EA must be concatenated; as was done for the analyses of section 4.2.1. In the 1991 Census survey design, extra stages of sampling were eliminated in most of the country and EA's generally correspond to PSU's. EA of residence for the household, as of December 31 of the reference year, is collected every year during the life of the panel. As households and individuals can have multiple EA's over the years, the same comments as the ones regarding multiple households of the previous paragraph also apply here. A unique EA was assigned to each individual using the same method as the one used for assigning unique household ID. This resulted in the 16,682 jobless spells being divided into 5,191 different EA's.

Immigrants behave differently from the rest of the Canadian population and are not included in any of the analyses of section 4.2.1. In SLID, immigrants are defined as individuals that were not born in Canada. They were easily identified using a summary flag from the SLID dataset which indicates if the individual is an immigrant or not. Another group of individuals excluded from the analyses of section 4.2.1, are people suffering from long-term disabilities, handicaps, or from any long-term physical or mental conditions. These individuals were identified using two summary flags from the SLID dataset: 1) long-term condition and 2) long-term disability. We combined these two variables into a more general "long-term health condition", which was then used to identified people who were just too ill to work and excluded them from all analyses.

As in SIPP, there is a seam-effect in the duration of jobless spells in SLID. However, the effect is considerably less as can be seen from figure 4.17. This is mainly the result of computer assisted interviews (CAI); see section 1.1.1. Hence, no correction was made for the analyses of section 4.2.1 and the computation of the baseline hazard functions of figures 4.17. The seam-effect is discussed further in Cotton & Giles (1998). Their figure 2, which shows the number of job starts by month, and figure 3, which shows the number of job ends by month, are particularly interesting. Figure 2 shows that May and June have the largest numbers of job starts; reflecting the importance of students and seasonal workers. There is also an unusually high number of job starts in January 1994, which is due to problems with the CAI system during the January 1995 interviews. Figure 3 shows that August, September and December have the largest numbers of job ends; reflecting the return to school for students in September and the end of Christmas jobs in December.

The number of jobless spells starting in each of the 6 years of the 1$^{\text{st}}$ panel are shown in table 4.5. The first column contains only the spells experienced by individuals with strictly positive longitudinal sampling weights and the second column contains all spells experienced by individuals from SLID 1$^{\text{st}}$ panel. Jobless spells are spread out fairly evenly over the 6 year period. Note that the differences between the values of the last row of table 4.5 and the sample sizes given in table 4.6 are due to missing values.

4.2.1 Jobless spells

As in section 4.1 and according to the discussion of section 3.3.2, three analyses based on model (3.2) were performed in section 4.2.1. These analyses are the same as the ones of sections 4.1.1 and 4.1.2; that is, the first two are weighted and unweighted analyses based on individuals with strictly positive sampling weights only and the last one is an unweighted analysis based on all individuals. The results are shown in table 4.6, with

year	spells with strictly positive weights only	all spells
1993	2,674	2,687
1994	2,781	3,093
1995	2,927	3,401
1996	2,861	3,458
1997	2,841	3,460
1998	2,808	3,312
Total	16,892	19,411

Table 4.5: Number of jobless spells starting each year.

standard deviations given in parentheses and where the variable Age^2 is the square of the Age variable.

The three columns of table 4.6 are the same analyses as the ones of tables 4.1 and 4.3, but applied to the SLID dataset with the household ID variable as clusters and provinces instead of states as strata. Column 1 contains the results of a weighted stratified marginal analysis based on 14,978 uncensored and right-censored jobless spells experienced by individuals with strictly positive longitudinal sampling weights only. This analysis uses the design-based methodology of Lin (2000), section 3, for super-population inference. The analysis in column 2 is the unweighted version of the one in column 1, and was done using the methodology proposed in chapter 3. In particular, the standard deviations were computed using $\widehat{V}_R(\hat{\beta})$, given by (3.19). In column 3, the sample of the analysis of column 2 is expanded to include all individuals; it is based on 16,682 uncensored and right-censored spells. The same comments regarding the strong similarities between the point estimates for β_0 for the three columns as the ones made after tables 4.1 and 4.3 also apply here. As the standard deviation estimates computed using EA as clusters are shown in table 4.7, more on this topic will follow that table.

Contrary to the tables presented in section 4.1, the real gain in terms of smaller

$\hat{\boldsymbol{\beta}}$	**weighted**	**unweighted**	
	$n = 14,978$	$n = 14,978$	$n = 16,682$
I(Sex)	-4.39×10^{-1} (1.61×10^{-1})	-4.33×10^{-1} (1.01×10^{-1})	-4.75×10^{-1} (9.56×10^{-2})
I(Looked)	6.41×10^{-2} (3.65×10^{-2})	3.42×10^{-2} (2.50×10^{-2})	4.71×10^{-2} (2.37×10^{-2})
I(Children)	-6.00×10^{-2} (4.33×10^{-2})	-6.64×10^{-2} (2.71×10^{-2})	-7.98×10^{-2} (2.56×10^{-2})
Age	-1.75×10^{-2} (3.64×10^{-3})	-1.12×10^{-2} (2.47×10^{-3})	-1.20×10^{-2} (2.34×10^{-3})
Education	-1.55×10^{-2} (7.95×10^{-3})	-1.04×10^{-2} (4.91×10^{-3})	-9.56×10^{-3} (4.64×10^{-3})
Hourly-wage	1.32×10^{-2} (3.28×10^{-3})	1.06×10^{-2} (2.44×10^{-3})	9.39×10^{-3} (2.32×10^{-3})
I(EI)	9.59×10^{-2} (3.93×10^{-2})	1.11×10^{-1} (2.51×10^{-2})	1.14×10^{-1} (2.34×10^{-2})
I(Minority)	-1.26×10^{-1} (1.57×10^{-1})	-1.87×10^{-1} (9.64×10^{-2})	-1.91×10^{-1} (8.79×10^{-2})
I(Winter)	7.72×10^{-1} (5.84×10^{-2})	8.12×10^{-1} (3.50×10^{-2})	8.01×10^{-1} (3.29×10^{-2})
Age^2	-1.09×10^{-3} (1.37×10^{-4})	-1.11×10^{-3} (9.13×10^{-5})	-1.07×10^{-3} (8.54×10^{-5})
I(Winter) $\times$ I(Weeks > 52)	-1.95 (5.90×10^{-2})	-1.98 (3.83×10^{-2})	-1.97 (3.62×10^{-2})
I(Sex)$\times$Education	4.03×10^{-2} (1.21×10^{-2})	3.43×10^{-2} (7.38×10^{-3})	3.61×10^{-2} (6.96×10^{-3})
I(Sex) $\times$ I(Children)	-1.34×10^{-1} (6.33×10^{-2})	-1.26×10^{-1} (4.04×10^{-2})	-1.31×10^{-1} (3.77×10^{-2})
Age$\times I$(Children)	1.30×10^{-2} (3.01×10^{-3})	1.17×10^{-2} (1.99×10^{-3})	1.28×10^{-2} (1.87×10^{-3})
I(Sex)$\times$Hourly-wage	-1.97×10^{-2} (5.61×10^{-3})	-1.81×10^{-2} (3.62×10^{-3})	-1.59×10^{-2} (3.48×10^{-3})
I(Winter) $\times$ I(Looked)	-1.58×10^{-1} (6.30×10^{-2})	-1.71×10^{-1} (3.98×10^{-2})	-1.88×10^{-1} (3.76×10^{-2})
Age$\times$Hourly-wage	-1.48×10^{-4} (1.96×10^{-4})	-5.69×10^{-4} (1.52×10^{-4})	-5.72×10^{-4} (1.44×10^{-4})
Age$\times I$(EI)	1.59×10^{-2} (2.81×10^{-3})	1.84×10^{-2} (1.97×10^{-3})	1.72×10^{-2} (1.87×10^{-3})

Table 4.6: Weighted and unweighted marginal Cox analyses for jobless spells.

standard deviations is the result that unweighted procedures are more efficient than weighted ones when the sampling design is uninformative (e.g., s.d.'s in column 2 are much smaller than the ones in column 1). As the importance of top-up samples is less in SLID than in SIPP, the ability to include all individuals, when using the proposed unweighted methods of section 3.4.2, does not yield substantial gain in terms of smaller standard deviations (e.g., s.d.'s in column 3 are slightly smaller than the ones in column 2). Both of these statements hold regardless of the clustering variable used: household ID or EA.

The model of table 4.6 is more complex (i.e., includes more main effects and interactions) than both models of tables 4.1 and 4.3. The main reasons behind this more complex model were to satisfy the assumption of uninformative sampling design and to yield a proper model for analytical inference on the duration of jobless spells. The model of table 4.6 is an improvement over the model of table A.1 (see appendix A.1), which was itself the result of improvements on primary SLID data analyses. In figures 4.11 and 4.12, martingale residuals, as defined by (4.2) and computed based on the model of table A.1, were plotted against age and education respectively. For confidentiality reasons, only the smooth regression curves (obtained from the S-Plus function loess.smooth) are shown and not the actual data points. Figure 4.11 suggests that the relation with age is quadratic and an age square term was added, which is highly statistically significant. This is not surprising as it is well known that the job market is harder for people in their early twenties as well as people in their fifties or sixties, than for people in their thirties or forties. Like figure 4.1, figure 4.12 suggests introducing a cut off point at approximately 15 years of education to allow for a different relationship between duration of jobless spells and education for individuals with more than 15 years of education. The term Education$\times I(\text{Education} > 15)$ was added to the model, but was not significant. The curve also has another critical point at approximately 10 years. However, there are few individuals with 10 or less years of education (data points not shown for confidentiality reasons) and adding an extra term to the model did not yield anything statistically significant.

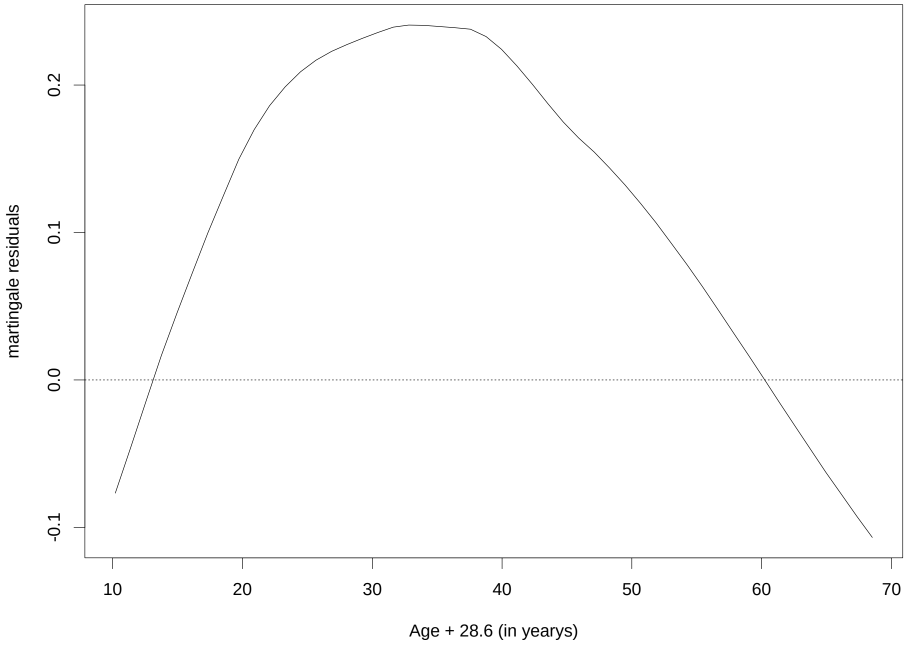

Figure 4.11: Martingale residuals vs. Age — model of table A.1.

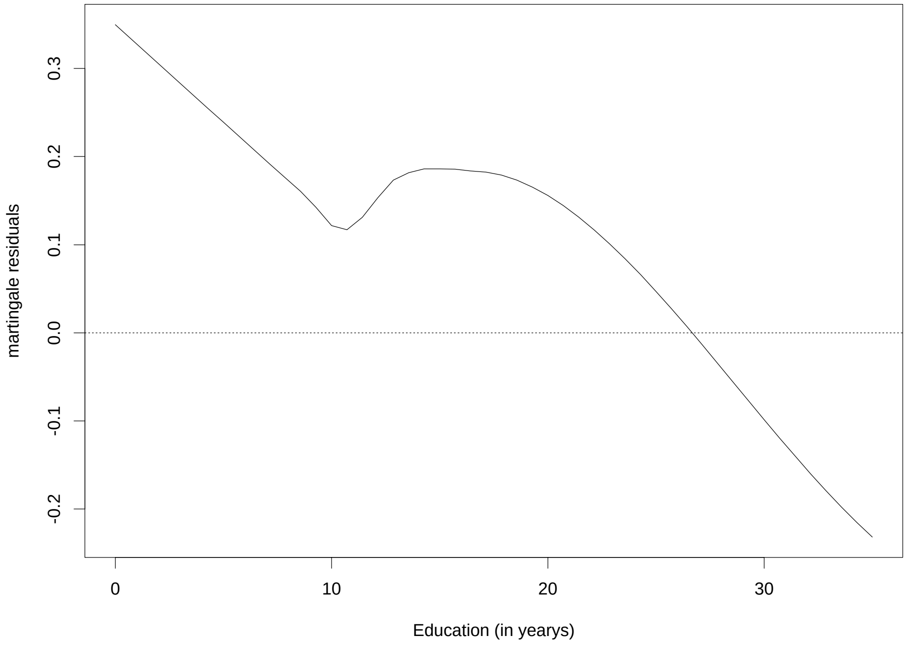

Figure 4.12: Martingale residuals vs. Education — model of table A.1.

Before going further with the analyses of the effects of the various covariates of table 4.6 on duration of jobless, it is preferable to look at the impact of choosing EA as the clustering variable instead of household ID. Table 4.7 shows the values of various standard deviation estimates, using both household ID and EA as the clustering variables. Those standard deviations include the ones of table 4.6. Since the choice of the clustering variable does not affect point estimates, they are not shown as they are identical to the ones given in table 4.6.

The first three columns of table 4.7 contain the standard deviation estimates computed using (3.19) with household ID as the clustering variable. Column 1 uses (3.75) and is based on 14,978 spells experienced by individuals with strictly positive longitudinal sampling weights only. Column 2 is the unweighted version of column 1, and column 3 is an extension of column 2 to include all 16,682 spells. The last three columns of table 4.7 show the results for the same weighted and unweighted estimators, but using EA as the clustering variable. The values for Lin's (2000) design-based standard deviation estimator for super-population inferences are not shown as they are practically identical to the the ones of column 1 and 4, depending on the clustering variable, to the accuracy shown in the table; similarly for Binder's (1992) estimates. The "naive" standard deviation estimates, which assume independence, are shown in table A.2 (see appendix A.1). They generally underestimate the standard deviations of the different covariates for the three models of table 4.6.

Table 4.7 allows comparisons between the different point estimates of table 4.6 using standard deviations with either household ID or EA as the clustering variable. Except for age, hourly-wage and the Age×Hourly-Wage interaction, the point estimates for the main effects and interactions of table 4.6 are not statistically different at the 10% level. This statement holds regardless of the variance estimator used to do the pairwise comparisons and regardless of whether household ID or EA is the clustering variable. The only two pairwise comparisons for hourly-wage that are not significant at the 10% level are between the models of columns 2 and 3 of table 4.6, when using $\widehat{V}_R(\hat{\beta})$ based on a sample

$\hat{\boldsymbol{\beta}}$	**Household ID as clusters**			**EA as clusters**		
	weighted	**unweighted**		**weighted**	**unweighted**	
	$n = 14,978$	$n = 14,978$	$n = 16,682$	$n = 14,978$	$n = 14,978$	$n = 16,682$
I(Sex)	1.61×10^{-1}	1.01×10^{-1}	9.56×10^{-2}	1.62×10^{-1}	1.06×10^{-1}	9.90×10^{-2}
I(Looked)	3.65×10^{-2}	2.50×10^{-2}	2.37×10^{-2}	3.75×10^{-2}	2.60×10^{-2}	2.45×10^{-2}
I(Children)	4.33×10^{-2}	2.71×10^{-2}	2.56×10^{-2}	4.17×10^{-2}	2.70×10^{-2}	2.52×10^{-2}
Age	3.63×10^{-3}	2.47×10^{-3}	2.34×10^{-3}	3.52×10^{-3}	2.49×10^{-3}	2.34×10^{-3}
Education	7.94×10^{-3}	4.91×10^{-3}	4.64×10^{-3}	7.86×10^{-3}	4.97×10^{-3}	4.66×10^{-3}
Hourly-wage	3.28×10^{-3}	2.44×10^{-3}	2.32×10^{-3}	3.26×10^{-3}	2.41×10^{-3}	2.30×10^{-3}
I(EI)	3.93×10^{-2}	2.51×10^{-2}	2.34×10^{-2}	3.92×10^{-2}	2.58×10^{-2}	2.37×10^{-2}
I(Minority)	1.57×10^{-1}	9.64×10^{-2}	8.79×10^{-2}	1.53×10^{-1}	9.25×10^{-2}	8.40×10^{-2}
I(Winter)	5.83×10^{-2}	3.50×10^{-2}	3.29×10^{-2}	5.65×10^{-2}	3.62×10^{-2}	3.43×10^{-2}
Age^2	1.37×10^{-4}	9.13×10^{-5}	8.54×10^{-5}	1.42×10^{-4}	9.47×10^{-5}	8.82×10^{-5}
I(Winter) $\times$ I(Weeks > 52)	5.88×10^{-2}	3.83×10^{-2}	3.62×10^{-2}	5.89×10^{-2}	3.81×10^{-2}	3.61×10^{-2}
I(Sex)$\times$Education	1.21×10^{-2}	7.38×10^{-3}	6.96×10^{-3}	1.20×10^{-2}	7.63×10^{-3}	7.14×10^{-3}
I(Sex) $\times$ I(Children)	6.33×10^{-2}	4.04×10^{-2}	3.77×10^{-2}	6.31×10^{-2}	4.10×10^{-2}	3.79×10^{-2}
Age$\times I$(Children)	3.01×10^{-3}	1.99×10^{-3}	1.87×10^{-3}	2.99×10^{-3}	2.03×10^{-3}	1.91×10^{-3}
I(Sex)$\times$Hourly-wage	5.61×10^{-3}	3.62×10^{-3}	3.48×10^{-3}	5.58×10^{-3}	3.63×10^{-3}	3.49×10^{-3}
I(Winter) $\times$ I(Looked)	6.29×10^{-2}	3.98×10^{-2}	3.76×10^{-2}	6.25×10^{-2}	4.21×10^{-2}	4.00×10^{-2}
Age$\times$Hourly-wage	1.96×10^{-4}	1.52×10^{-4}	1.44×10^{-4}	1.98×10^{-4}	1.51×10^{-4}	1.44×10^{-4}
Age$\times I$(EI)	2.81×10^{-3}	1.97×10^{-3}	1.87×10^{-3}	2.76×10^{-3}	1.94×10^{-3}	1.84×10^{-3}

Note: s.d.'s computed using $\widehat{V}_R(\hat{\beta})$, given by (3.19), and $\widehat{V}_{R_W}(\hat{\beta}_W)$, given by (3.75).

Table 4.7: Standard deviation estimates for some of the different weighted and unweighted marginal Cox PH analyses for jobless spells.

size of 16,682 jobless spells and for both clustering variables. Both of these pairwise comparisons are just under the 10% level. Similar comments also apply to age; except, that there are four pairwise comparisons which are below the 10% level. As for hourly-wage two of these pairwise comparisons (models of column 1 v.s. 2, using Lin's (2000) super-population variance estimator and for both clustering variables) are just under the 10% level, but the other two (models of column 2 v.s. 3, using $\widehat{V}_R(\hat{\beta})$ based on a sample size of 16,682 and for both clustering variables) are just above the 1% level. Eight of the twelve pairwise comparisons between point estimates for the Age×Hourly-Wage interaction are statistically significant at the 5% or 1% level. Although this could raise some doubts about the uninformative and/or ignorability assumptions as well as if individuals that joined existing households after the start of the panel in January 1993, these doubts are limited as this interaction term is not statistically significant in any of the three models of table 4.6; see table 4.8 and the following discussion for further information. Note that when using the Bonferroni correction for multiple comparisons, none of these pairwise comparisons are statistically significant at the 10% level. However, with 216 such comparisons, the validity of the Bonferroni correction may be questionable. In addition, the part of the discussion following table 4.1, in which it was mentioned that an exact pairwise test would involve a more complex variance term as its denominator, applies here as well. In summary, tables 4.6 and 4.7 imply that the assumptions of uninformative and/or ignorable sampling designs are satisfied. They also justify the use of model (3.2), with the covariates, clusters and strata described in section 4.2 and table 4.6, to analyse jobless spells. In addition, including the individuals that joined existing households after the start of the panel in the analysis of column 3, table 4.6, is supported by these pairwise comparisons between the point estimates of table 4.6.

Strong similarities between the first three columns and the, corresponding, last three columns of table 4.7, imply that using either household ID or EA (the PSU) as the clustering variable yield similar standard deviation estimates. Hence, both choices are valid for inference purposes. Unweighted standard deviation estimates of columns 2 and 5 are substantially smaller than the ones of columns 1 and 4, respectively. This illustrates the result that unweighted procedures are more efficient than weighted ones when the sampling design is uninformative and/or ignorable. Recall that this gain in efficiency was not as important in tables 4.1 and 4.3. The proportion of jobless spells experienced by individuals with zero sampling weights is much less important than for either spells on the food stamps program or without health insurance; see sections 4.1.1 and 4.1.2. For the food stamps program, top-up samples were comprised from a total of 721 spells, which represented 39.7% of all spells in the program. With a proportion of 43.7% of all spells without health insurance experienced by individuals with zero sampling weights, the prevalence of top-up sample spells is even higher in that case. Compared to these proportions, only 10.2% of the jobless spells are from individuals with zero sampling weights. This explains why the gain in terms of smaller standard deviations is not as important when comparing columns 3 and 6 to columns 2 and 5 in table 4.7

Summary of jobless spells analyses

Column 1 of table 4.8 contains the largest p-value of the three models of table 4.6 for each of the main effects and interactions, with household ID as clustering variable. Column 2 is analogous, but EA is the clustering variable instead of household ID. As with the previous tables, the values of the two columns of table 4.8 are very similar. This feature should not be surprising after the discussion of table 4.7 and its implications. To summarize table 4.8, except for looked, children, minority and the Age×Hourly-Wage interaction, all the main effects and interactions of table 4.6 are statistically significant at the 5% level[3]. Keeping in mind the previous remarks on education, this statement is true regardless of whether household ID or EA is the clustering variable. Looked and children were kept in the final

[3] Note that education is statistically significant at the 5% level when EA is the clustering variable, but just above that threshold when household ID is used instead. This p-values of 0.051 correspond to column 2 of table 4.6, which is based on a weighted model that is known to be less efficient. The largest p-value for education between columns 1 and 3 is 0.039. Hence, it is reasonable to conclude that education is significant at the 5% level.

$\hat{\boldsymbol{\beta}}$	**Clusters**	
	Household ID	**EA**
I(Sex)	6.5×10^{-3}	6.8×10^{-3}
I(Looked)	1.7×10^{-1}	1.9×10^{-1}
I(Children)	1.7×10^{-1}	1.5×10^{-1}
Age	5.9×10^{-6}	7.2×10^{-6}
Education	5.1×10^{-2}	4.8×10^{-2}
Hourly-wage	5.4×10^{-5}	4.9×10^{-5}
I(EI)	1.5×10^{-2}	1.5×10^{-2}
I(Minority)	4.2×10^{-1}	4.1×10^{-1}
I(Winter)	0.0	0.0
Age^2	1.7×10^{-15}	2.0×10^{-14}
$I(\text{Winter}) \times I(\text{Weeks} > 52)$	0.0	0.0
I(Sex)×Education	8.2×10^{-4}	8.1×10^{-4}
$I(\text{Sex}) \times I(\text{Children})$	3.4×10^{-2}	3.3×10^{-2}
Age×I(Children)	1.4×10^{-5}	1.3×10^{-5}
I(Sex)×Hourly-wage	4.3×10^{-4}	4.0×10^{-4}
$I(\text{Winter}) \times I(\text{Looked})$	1.2×10^{-2}	1.2×10^{-2}
Age×Hourly-wage	4.5×10^{-1}	4.5×10^{-1}
Age×I(EI)	1.5×10^{-8}	8.3×10^{-9}

Table 4.8: Largest p-values for the different weighted and unweighted analyses of table 4.7, with household ID or EA as the clustering variable.

model as interaction terms involving these two variable (i.e., $I(\text{Winter}) \times I(\text{Looked})$ for looked, and $I(\text{Sex}) \times I(\text{Children})$ and Age×I(Children) for children) are significant. The reasons for keeping minority and the Age×Hourly-Wage interaction in the model are more open to discussion. First, it is logical that both of them should influence the duration of jobless spells. Second, and more important, including these two terms improves the model by making it uninformative and/or ignorable, which is required to make analytical inference and conclusions. This second reason also applies to looked and children.

The following conclusions on the effects of the various covariates can be reached from tables 4.6 to 4.8:

- Women are more likely to experience longer jobless spells;
- This propensity towards longer jobless spells for women is heightened if she has one or more children;
- For men, having one or more children has no effect on duration of jobless spells;
- For individuals younger than 22.95 years, the younger you are the more likely you are to experience longer jobless spells, and the quadratic term implies that this is worse for the youngest (recall that only individuals aged 16 years and over can have jobless spells in SLID);
- For individuals older than 22.95 years, the older you are the more likely you are to experience longer jobless spells, and the quadratic term implies that this is worse for the oldest;
- For men, more education implies that they are more likely to experience longer jobless spells (a surprising result, which is probably balanced by men being less likely to become jobless if they are more educated);
- For women, more education implies that they are more likely to experience shorter jobless spells;
- For men, a higher hourly-wage implies that they are more likely to experience shorter jobless spells;
- For women, a higher hourly-wage implies that they are more likely to experience longer jobless spells (although, the effect is marginal);

- Individuals receiving EI are more likely to experience longer jobless spells;
- Individuals who lost their job in the winter (as defined earlier) are more likely to experience shorter jobless spells (this confirms the effect of seasonal jobs[4]);
- However, winter has the opposite effect if the individual has been without work for more than 52 weeks.

Diagnostic and model checking

The same comments regarding diagnostic and model checking for unweighted analyses made at the end of sections 4.1.1 and 4.1.2 also apply here. The "naive" standard deviation estimates, which assume independence, for the unweighted models of table 4.6 are generally smaller than their corresponding unweighted standard deviations given in tables 4.6 and 4.7, but the differences are not substantial. Their values are shown in table A.2, when household ID is the clustering variable (see appendix A.1). This implies that standard model checking procedures, which do not account for correlation, should not be too affected by the effect of correlated jobless spells. Figures 4.13 to 4.16, which test the proportional hazards assumption using the method introduced by Grambsch & Therneau (1994), did not reveal any serious departure from model (3.2). However, figures 4.15 and 4.16 show a small variation over time of the values of $\hat{\beta}$ for I(Looked) and $I(\text{Winter}) \times I(\text{Looked})$. Other graphics and diagnostic procedures did not reveal any serious departure from model (3.2). Note that many diagnostic plots and procedures can not be shown for confidentiality reasons.

Testing the ignorability assumption of the sampling design using (3.68) was not done, as this test has very low power; see section 3.3.1. Instead, the pairwise comparisons described previously were preferred. Again, there is no indication that the uninformative

[4] Another confirmation of the effect of seasonal jobs is given by the fact that individuals who lost their job in the winter and were looking for work, which probably imply that they are not seasonal workers, are more likely to experience longer jobless spells.

and/or ignorable assumptions are not satisfied.

Finally, one could argue in favor of only including the jobless spells that started in 1993 in the analyses of section 4.2.1. This would result in the exclusion of all jobless spells that started in 1994–1998 (inclusively) and has yielded a sample size of 2,313 spells compared to the 14,978 spells of table 4.6 (or 16,682 when including top-up samples). The reasoning behind this choice of only including the jobless spells that started in 1993 is due to the dynamic nature of the population; see section 2.5. Consequently, jobless spells that started in 1997 may behave much differently than those that started in 1993 (e.g., this would be the case if an economic recession had started in 1995). Lawless & Boudreau (2002), table 1, contains such a weighted and an unweighted analyses for jobless spells that started in 1993 only. They also used a stratified Cox model and their models include most of the covariates found in table 4.6. Except of $I(\text{EI})$ and $I(\text{Winter})$, their fitted models are in fairly good agreement with the ones of table 4.6, columns 1 and 2.

Cumulative baseline hazard functions

The cumulative baseline hazard function $\Lambda_{0h}(t, \beta)$ for the duration of jobless spells in a given province (i.e., stratum) can be estimated using the Breslow-Aalen estimator, given by (3.34). Figure 4.17 shows these estimated functions for individuals who experienced jobless spells in Ontario (solid line) or in Newfoundland (dotted line). These estimated cumulative baseline hazard functions are based on the model of table 4.6, column 2. The estimated curves of figure 4.17 confirm the general notion that individuals living in Newfoundland are more likely to experience longer jobless spells than individuals living in Ontario.

Chapters 3 and 4 were mainly concerned with variance estimation procedures that account for the use of complex survey designs, with stratification and clustering, in the collection of longitudinal survey data. Another particularity of such surveys is the presence of left-truncated observations; see sections 1.4 and 2.2. In chapter 5, the conditional

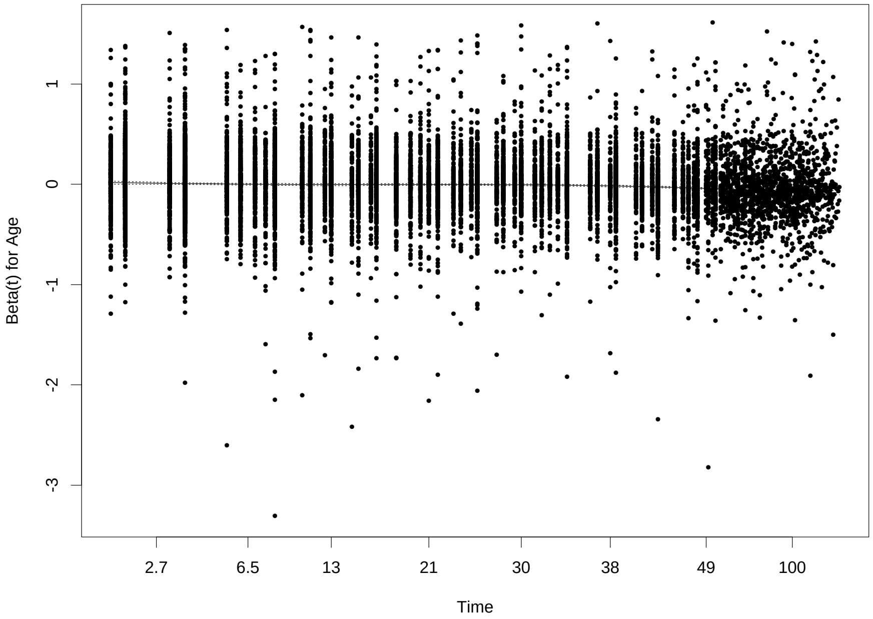

Figure 4.13: Test of PH assumption for Age in model of table 4.6, column 2.

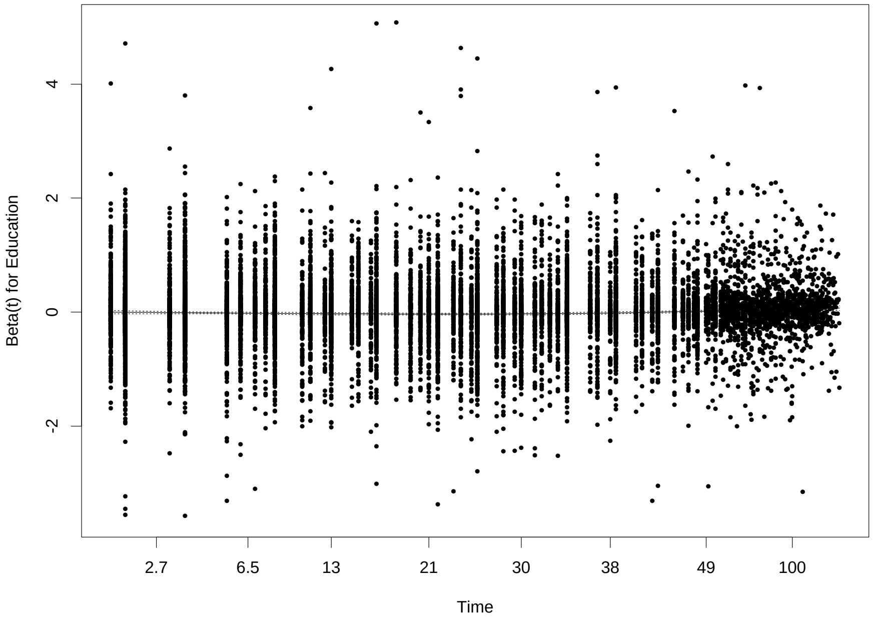

Figure 4.14: Test of PH assumption for Education in model of table 4.6, column 2.

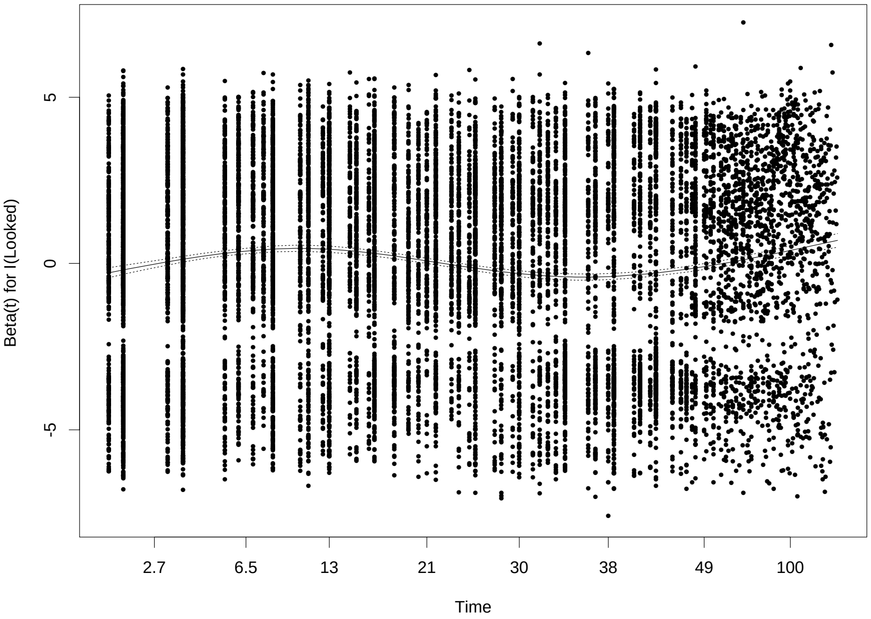

Figure 4.15: Test of PH assumption for I(Looked) in model of table 4.6, column 2.

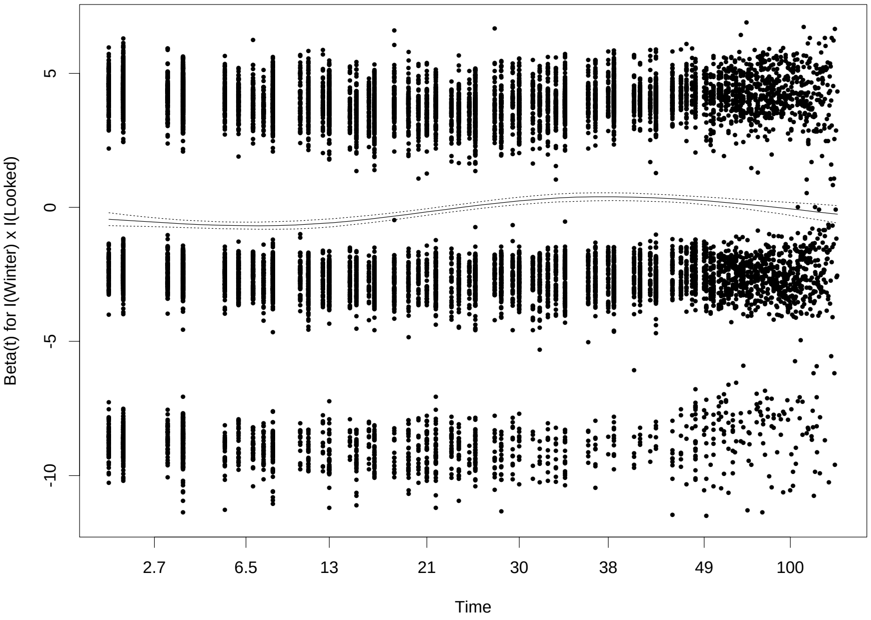

Figure 4.16: Test of PH assumption for I(Winter) × I(Looked) in model of table 4.6, column 2.

partial likelihood, a generalization of the partial likelihood given by (3.6) that allows the inclusion of left-truncated sojourns, is discussed. Other methods for including left-truncated observations are also discussed and compared in the next chapter.

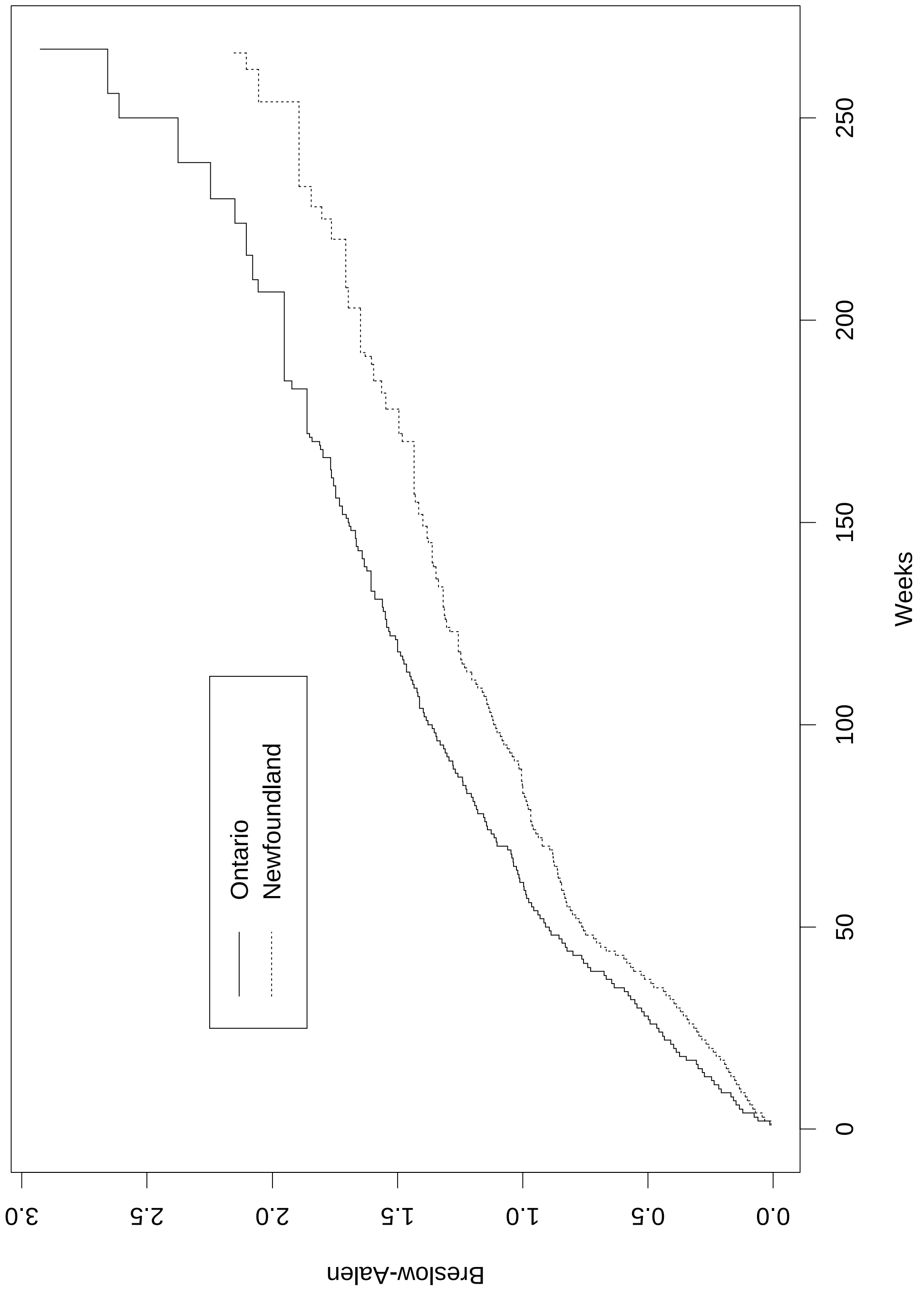

Figure 4.17: Estimated cumulative baseline hazard functions for individuals who experienced jobless spells in Ontario or in Newfoundland.

Chapter 5

Left-truncation

Individuals that are selected to be part of a longitudinal survey and that were at risk of experiencing the event or transition of interest for some time period before the start of the observation period are said to be left-truncated, if that event or transition did not occur prior to the start of the study. This implies that time intervals actually observed for these left-truncated individuals tend to be longer than those arising from the true underlying failure distribution. This was first discussed in section 1.4 and portrayed in figure 1.9. In addition, the prevalence of left-truncated sojourns in SIPP and PSID was illustrated in section 2.2. For example, 49.5% of individuals that benefited from food stamps were already on the program before the start of the 1987 SIPP panel (see table 2.1). Similarly, about 45% of jobless spells in SLID 1st panel had begun prior to January 1, 1993. Hence, the individuals experiencing these spells were at risk of finding a job before 1993 and, if it had been the case, the jobless spell in question would not have been recorded by SLID interviewers.

Chapter 5 builds on section 1.4 to explore different methods for handling left-truncated sojourns. In section 5.1, the conditional Cox partial-likelihood method, which can handle left-truncated sojourns when the lengths of the exposure periods prior to the start of the observation window are known, is discussed. This method is an extension of the Cox partial-likelihood used in chapters 3 and 4, and has many potential applications in longitudinal surveys. In section 5.2, three methods or scenarios associated with left-truncated observations are presented. They consist of: 1 — the conditional approach (section 5.2.1); 2 — using the backward recurrence time (section 5.2.2); and 3 — the equilibrium assumption (section 5.2.3). In section 5.3, these three scenarios are discussed and compared in the context of the Exponential distribution. In section 5.4, the same three scenarios are considered in the context of the Weibull distribution. Section 5.4.1 is limited to the case when there is no censored observation. In that case it is possible to analytically derive results regarding comparisons between the Fisher information matrices of the three scenarios of section 5.2. Unfortunately, the complexity of the formulas when censoring is allowed (see section 5.4.2), makes such comparisons only possible through simulations; this is done in section 5.4.3.

As mentioned previously, Guo (1993) contains an interesting overview on left-truncation and its numerous complications on statistical inference. He also gives various methods to deal with left-truncated observations including some that are very similar to those presented in sections 5.1 and 5.2. The presence of left-truncated observations is fairly common to econometrists, and Heckman & Singer (1986), section 1.5 in particular, present a detailed theoretical discussion on left-truncated observations and their related statistical distributions for various cases. Finally, Asgharian et al. (2002), whose work is partially based on Vardi (1989), compared conditional and unconditional approaches for estimation of length-biased observations. This has connections with the comparisons of sections 5.3 and 5.4.

5.1 Conditional partial-likelihood

Many longitudinal surveys (e.g., SLID, SIPP and PSID) focus on social and economic events experienced by sampled individuals during part of their life, which corresponds to the life of the panel. Studying these types of events (e.g., duration of jobless spells, duration of participation to different social programs and duration of marital unions) has the advantage that it is often possible to assess retrospectively the start of the exposure period if it falls before the start of the observation window; see section 1.2.4. In other words, b_{hi}, the exposure time prior to the start of the study for the i^{th} individual of the h^{th} stratum, is determined retrospectively for left-truncated individuals or sojourns. Even though these left-truncated sojourns were not used in the analyses of section 4.2, a large proportion of the start dates for jobless spells that started prior to January 1, 1993 in SLID 1^{st} panel (see section 1.1.1) were assessed retrospectively and were available in the dataset. Similarly, the start date of marital unions in PSID (see section 1.1.3) are available even if the marriage began prior to 1968, the start of the study. However, this is not the case for spells on the various social programs and periods without health insurance studied in SIPP (see section 1.1.2). For these spells it is only known that they are left-truncated. Note that it is much harder to assess retrospectively the start of the exposure period for things like cancer or HIV infection. In other words, longitudinal surveys are fortunate as human subjects are much more likely to know and remember the dates of life events than the dates of other types of events. Note that whether the b_{hi}'s are known or not, discarding left-truncated sojourns or observations does not create bias; see Allison (1984), for example. However, as discussed in section 2.2, this might result in substantial reduction of the sample size and in the loss of information on observations with unique characteristics.

Obviously, this raises valid questions regarding possible recall bias. In addition, the usual comments on the quality of retrospective data apply here as well. The seam-effect discussed in section 2.1 clearly illustrates that human memory is not perfect. Depending on the life history events being studied, recall bias problems maybe more or less of a concern. For example, when studying duration of marital unions, one would expect that recall bias should be negligible. Freedman et al. (1988) reported a remarkable correspondence between dates collected prospectively as part of the ongoing PSID longitudinal survey and the same dates assessed retrospectively years afterwards.

If the start of the exposure period can be assessed retrospectively when it falls before the start of the observation window, conditional methods as described in section 1.4.1 can be used. The idea is to account for the bias created by the selection of left-truncated observations in the sample by conditioning on the known values of b_{hi}'s, which were assessed retrospectively. In the case of the stratified Cox model of (3.2), the partial-likelihood function given by (3.6) is now referred to as a *conditional partial-likelihood*, and takes the following form

$$L(\beta) = \prod_{h=1}^{H} \prod_{i=1}^{n_h} \left(\frac{\exp\{\beta' x_{hi}(t_{hi})\}}{\sum_{j=1}^{n_h} \gamma_{hj}^{b}(t_{hi}) \, \exp\{\beta' x_{hj}(t_{hi})\}} \right)^{\delta_{hi}} , \tag{5.1}$$

where $\gamma_{hi}^{b}(t) = I(b_{hi} \le t \le t_{hi}) = I(t_{hi} \ge t) \times I(b_{hi} \le t)$ and b_{hi} is the length of time that has elapsed before the start of the observation window; see sections 1.4.1 and 5.2.1. The difference between (3.6) and (5.1) is that $\gamma_{hi}(t)$ in the former is replaced by $\gamma_{hi}^{b}(t)$ in the latter. Therefore, the denominator in (5.1) is restricted by the additional condition in $\gamma_{hi}^{b}(t)$; that is, the i^{th} individual of the h^{th} stratum must have started its observation period at time t (i.e., $b_{hi} \le t$). Note that for individuals that are not left-truncated $b_{hi} = 0$ and $\gamma_{hi}^{b}(t) = \gamma_{hi}(t)$, as defined by (3.3). For (5.1) to be a proper conditional partial-likelihood, the independent left-truncation assumption of section 1.4.1 must be satisfied $\forall\,(h,i)$'s. This assumption, given by (1.29), only acquires its full meaning when covariates are introduced in the model. In that case, the transition time T and the left-truncation time (or entry time) L are conditionally independent, given the covariate $x(t)$; that is,

$$\Pr\{t \le T < t + \Delta t \mid T \ge t, L, T > L, x(t)\} = \Pr\{t \le T < t + \Delta t \mid T \ge t, x(t)\} \; ; \tag{5.2}$$

see Lawless (2003), section 2.4.1. Conditional independent left-truncation implies that, by defining the risk set carefully, the Cox model is still valid. However, as mentioned by Cnaan & Ryan (1989), problems where $x(t)$ involves time varying covariates require extra attention. In addition to (5.2), the definitions of uninformative and ignorable sampling designs (see definitions 3.1 and 3.2, respectively), must be modified to include the b_{hi}'s.

This situation described in this section is portrayed in figure 5.1. Guo (1993), sections 3–6, contains a good overview and an example of conditional partial-likelihood methods. In figure 5.1, the risk set at time t_1 is composed of individuals #1–4. Individual #5 is not part of the risk set at time t_1 since $b_5 > t_1$, which implies that this individual was not under study at time t_1. However, $b_4 < t_1$ and individual #4 is part of the risk set at time t_1. Similarly, the risk set at time t_2 is composed of individuals #2–5.

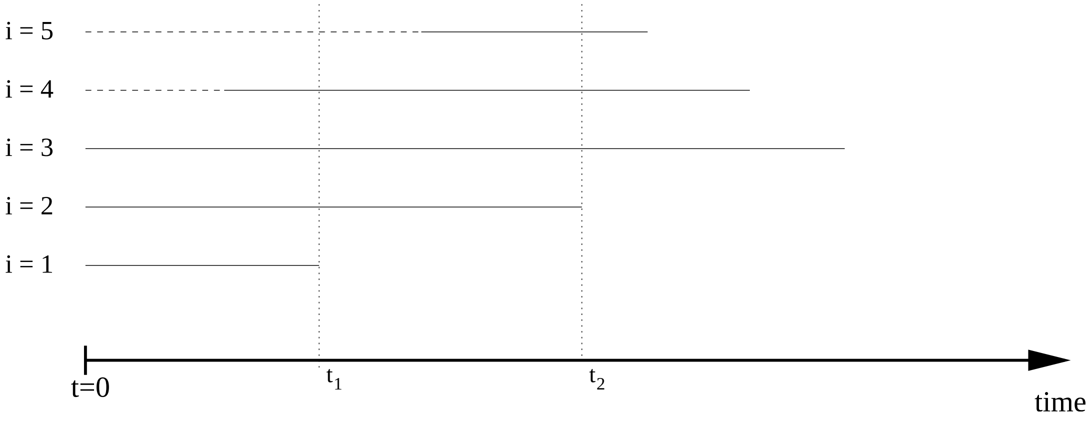

Figure 5.1: Risk set in the conditional partial-likelihood.

As mentioned in the first paragraph of this section, the values of the b_{hi}'s may not always be known or available in longitudinal survey analysis. In that case, it is much harder to include left-truncated sojourns into the analyses, especially with covariates which might depend on the unknown b_{hi}'s. Strong assumptions, like the equilibrium assumption of section 5.2.3, are required to draw any meaningful inference.

In addition, some surveys have addition information regarding the b_{hi}'s. For example, it might be possible to estimate their distribution, the so-called entry rate function. If parameters of this entry rate function involve or depend on the parameter of the distribution of the t_{hi}'s, it likely contains valuable information for inference on sojourn duration inference. This is particularly true if the entry rate distribution corresponds to the backward recurrence distribution that will be discussed further in section 5.2.2.

The questions raised in the last two paragraphs are explored in greater details for simpler models than the stratified Cox model; that is: the Exponential distribution in section 5.3 and the Weibull distribution in section 5.4. The general setting is first given in section 5.2. However, Cox model methodology that can handle unknown values of b_{hi}'s is an area where research is needed.

5.2 Three scenarios

Building on section 1.4, let $T_1^*, \ldots, T_n^*$ be i.i.d. random lifetime variables (e.g., Exponential, Weibull, Gamma), with p.d.f. $f(t)$, c.d.f. $F(t)$, survivor function $S(t)$ and mean $E(T_i^*) = \mu_T$. As discussed at the beginning of chapter 5, if t_i corresponds to a left-truncated time, its p.d.f. is not $f(t)$, but rather the matching length-biased p.d.f. $f_{LB}(t)$ defined as

$$f_{LB}(t) = \frac{t}{\mu_T} f(t) \quad \text{for } t \geq 0 \;, \tag{5.3}$$

under the equilibrium assumption. Note that (5.3) is also valid under the assumption of constant entry rate; see section 1.4. The length-biased distribution accounts for the fact that the i^{th} observation was at risk of experiencing the event before the start of the study. Hence, time intervals from the length-biased p.d.f. $f_{LB}(t)$ are longer than the corresponding ones from the true underlying failure p.d.f. $f(t)$; that is, $S_{LB}(t) \geq S(t)$, which is known as the *inspection paradox*.

Let T_i be a length-biased lifetime random variable; it can be re-expressed as: $T_i = B_i + \widetilde{T}_i$, where B_i is called the *age* at the time of selection and $\widetilde{T}_i$ is called the *excess* or *residual life*; see Ross (1996), section 3.4.1. This situation is portrayed in figure 5.2. In other words, the random variable B_i is the backward recurrence time corresponding to T_i and the random variable $\widetilde{T}_i$ is the forward recurrence time. Their p.d.f.'s are defined as

$$f_B(t) = f_{\widetilde{T}}(t) = \frac{1}{\mu_T} S(t) \quad \text{for } t \geq 0 \;, \tag{5.4}$$

under the equilibrium assumption.

In section 5.2.3, where it is assumed that the process is in equilibrium, it is only necessary to know $\tilde{t}_1, \ldots, \tilde{t}_n$ and inference is based solely on the forward recurrence p.d.f.

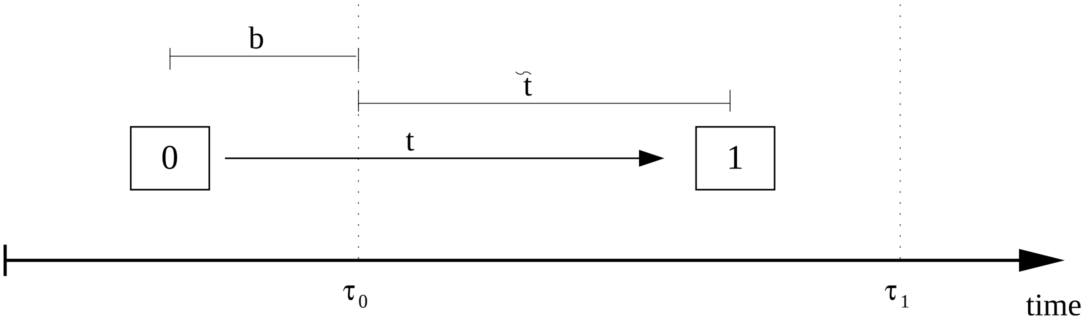

Figure 5.2: Length-biased, age and residual life variables.

$f_{\tilde{T}}(t)$ defined by (5.4); see Ross (1996), page 131. This is not the case in sections 5.2.1 and 5.2.2, where knowledge of both $\tilde{t}_1, \ldots, \tilde{t}_n$ and $b_1, \ldots, b_n$ is required to carry statistical inference. In section 5.2.2, it is further assumed that the entry rate $g(b)$ is the equilibrium backward recurrence distribution corresponding to T_i and, thus, is given by (5.4).

Classic references on delayed entry, another term for left-truncation, and their related processes (e.g., backward recurrence time and equilibrium distribution) include: Grimmett & Stirzaker (1992) and Ross (1996).

5.2.1 Conditional

For the first scenario, let $(\tilde{t}_1, \delta_1), \ldots, (\tilde{t}_n, \delta_n)$ be observed left-truncated failure or censored times, with origin at the beginning of the study, and censoring indicators δ_i's. In addition, it is assumed that $b_1, \ldots, b_n$ are known. They are either accessed retrospectively or by other means. The likelihood function corresponding to $(\tilde{t}_1, \delta_1), \ldots, (\tilde{t}_n, \delta_n)$, conditional on the known values of $b_1, \ldots, b_n$, is given by

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \Pr\{\tilde{t}_i, \delta_i \mid b_i, t_i \geq b_i\} \\ &= \prod_{i=1}^{n} \frac{f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}}{S(b_i)} , \qquad (5.5) \end{aligned}$$

where $t_i = \tilde{t}_i + b_i$ by definition and $\delta_i = I(t_i$ corresponds to a failure time).

5.2.2 Backward recurrence time

For the second scenario, the assumptions of section 5.2.2 are similar those of section 5.2.1. However, in addition to observing $(\tilde{t}_1, \delta_1), \ldots, (\tilde{t}_n, \delta_n)$ and knowing the values of $b_1, \ldots, b_n$, it is also known or assumed that $B_i \sim f_B(b)$, for $i = 1, \ldots, n$. Including that extra information in the likelihood function yields the following

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \Pr\{\tilde{t}_i, b_i, \delta_i \mid t_i \geq b_i\} \\ &= \prod_{i=1}^{n} f_B(b_i) \, \frac{f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}}{S(b_i)} \ , \end{aligned} \tag{5.6}$$

where $f_B(b)$ is given by (5.4) and where t_i and δ_i were defined in (5.5).

5.2.3 Equilibrium

For the third scenario, knowledge of the b_i's is not required in section 5.2.3 and only $(\tilde{t}_1, \delta_1), \ldots, (\tilde{t}_n, \delta_n)$ need to be observed. To achieve this it is assumed that the process is in equilibrium or, equivalently, that the rate of occurrence of the start of spells prior to τ_0 is constant. This is a fairly strong assumption which is generally difficult or impossible to test. Under this assumption, the likelihood is a function of (5.4) and is given by,

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \Pr\{\tilde{t}_i, \delta_i \mid \text{the process is in equilibrium}\} \\ &= \prod_{i=1}^{n} \left(\frac{1}{\mu_T} S(\tilde{t}_i) \right)^{\delta_i} \left(\int_{\tilde{t}_i}^{\infty} \frac{1}{\mu_T} S(u) \, du \right)^{1-\delta_i} \ . \end{aligned} \tag{5.7}$$

Note that there are other methods that allow the inclusion of left-truncated observations where the values of the b_i's are unknown. For example, the method proposed by Kalton et al. (1992), chapter 4, for the Kaplan-Meier estimator includes the left-truncated sojourns in the risk set only. Unfortunately, this modified Kaplan-Meier estimator has to be made unbiased by multiplying by a term, which is derived under fairly strong stationarity assumptions.

5.3 Exponential

Section 5.3 explores some of the points made by Guo (1993) and in section 1.4.2 in greater detail for the Exponential distribution. This section has some similarities with the work of Cox (1969) on the estimation of fiber lengths. The memoryless property of the Exponential makes derivations and comparisons of the three scenarios of section 5.2 simpler. Hence, section 5.3 will also serve as a building block for section 5.4, which contains the new and original work.

Let $T_1, \ldots, T_n$ be i.i.d. Exp(λ); that is, follow an Exponential distribution with scale parameter λ. The Exponential p.d.f. is,

$$f(t) = \lambda \exp\{-\lambda t\} \quad \text{for } t \geq 0 \; , \tag{5.8}$$

where $\lambda > 0$.

The memoryless property of the Exponential implies that $f(t) = f_B(t) = f_{\widetilde{T}}(t)$ and $f_{LB}(t) = \lambda^2 \, t \, \exp\{-\lambda t\}$; see (5.3) and (5.4).

5.3.1 Scenarios of section 5.2 for the Exponential

In section 5.3.1, the likelihood, log-likelihood, score, observed and Fisher information functions are computed for each of the scenarios of section 5.2 when $T_i \sim$ Exp(λ), for $i = 1, \ldots, n$. The information functions and variances will then be compared in section 5.3.2.

Conditional or scenario 1

For the Exponential distribution, with p.d.f. given by (5.8), (5.5) simplifies to yield,

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{\delta_i} \exp\{-\lambda t_i\}}{\exp\{-\lambda b_i\}} \; . \tag{5.9}$$

The corresponding log-likelihood, score, observed and Fisher information functions are,

$$l(\lambda) = d \ln \lambda - \lambda \sum_{i=1}^{n} \tilde{t}_i \qquad \text{since } t_i = b_i + \tilde{t}_i \ , \tag{5.10}$$

$$U(\lambda) = d/\lambda - \sum_{i=1}^{n} \tilde{t}_i \ , \tag{5.11}$$

$$\mathcal{I}(\lambda) = d/\lambda^2 \tag{5.12}$$

and

$$I(\lambda) = E(d)/\lambda^2 \ , \tag{5.13}$$

where $d = \sum_{i=1}^{n} \delta_i$.

Backward recurrence time or scenario 2

For the Exponential distribution, with p.d.f. given by (5.8), (5.6) simplifies to yield,

$$L(\lambda) = \prod_{i=1}^{n} \lambda \exp\{-\lambda b_i\} \, \frac{\lambda^{\delta_i} \exp\{-\lambda t_i\}}{\exp\{-\lambda b_i\}} \ . \tag{5.14}$$

The corresponding log-likelihood, score, observed and Fisher information functions are,

$$l(\lambda) = (n+d) \ln \lambda - \lambda \sum_{i=1}^{n} t_i \ , \tag{5.15}$$

$$U(\lambda) = (n+d)/\lambda - \sum_{i=1}^{n} t_i \ , \tag{5.16}$$

$$\mathcal{I}(\lambda) = (n+d)/\lambda^2 \tag{5.17}$$

and

$$I(\lambda) = (n + E(d))/\lambda^2 \ . \tag{5.18}$$

Equilibrium or scenario 3

For the Exponential distribution, with p.d.f. given by (5.8), (5.7) simplifies to yield,

$$L(\lambda) = \prod_{i=1}^{n} \lambda \exp\{-\lambda \tilde{t}_i\} \lambda^{\delta_i - 1} \ . \tag{5.19}$$

The corresponding log-likelihood, score, observed and Fisher information functions are,

$$l(\lambda) = d \ln \lambda - \lambda \sum_{i=1}^{n} \tilde{t}_i \ , \tag{5.20}$$

$$U(\lambda) = d/\lambda - \sum_{i=1}^{n} \tilde{t}_i \ , \tag{5.21}$$

$$\mathcal{I}(\lambda) = d/\lambda^2 \tag{5.22}$$

and

$$I(\lambda) = E(d)/\lambda^2 \ . \tag{5.23}$$

5.3.2 Comparisons

In this section, the results of section 5.3.1 are compared. Comparisons of the Fisher information functions $I(\lambda)$ yield that,

$$(5.13) = (5.23) = \frac{E(d)}{\lambda^2} < \frac{n + E(d)}{\lambda^2} = (5.18) \ . \tag{5.24}$$

That is, scenarios 1 and 3 contain the same information and scenario 2 has the most information. In addition, scenario 2 contains twice as much information as the other two scenarios when all the observations are failures (i.e., $d = n$). The comparison made in (5.24) can also be expressed in terms of $\mathrm{Var}(\hat{\lambda})$, where $\hat{\lambda}$ is the m.l.e. for each of the three scenarios. The relation $\mathrm{Var}(\hat{\lambda}) = \mathcal{I}^{-1}(\lambda)$ implies that,

$$\mathrm{Var}(\hat{\lambda}_2) = \frac{\lambda^2}{n + d} < \frac{\lambda^2}{d} = \mathrm{Var}(\hat{\lambda}_1) = \mathrm{Var}(\hat{\lambda}_3) \ , \tag{5.25}$$

where $\hat{\lambda}_j$ is the m.l.e. for the j^{th} scenario, $j = 1, 2, 3$. Hence, $\hat{\lambda}_1$ and $\hat{\lambda}_3$ have the same variance, which is larger than the variance of $\hat{\lambda}_2$. If $d = n$, $\text{Var}(\hat{\lambda}_2) = 1/2\,\text{Var}(\hat{\lambda}_1) = 1/2\,\text{Var}(\hat{\lambda}_3)$.

In addition to the analytical results given in (5.24) and (5.25), a small simulation study was done. It involved two sample sizes ($n = 10$ and $n = 100$) and 20 values of the scale parameter λ ($\lambda = 0.1, 0.2, \ldots, 2$). For each of these 40 combinations of (n, λ), n observations from the backward recurrence distribution were randomly generated along with another n random observations from the forward recurrence distribution. Each of the n length-biased sojourns was then computed by summing the backward and forward recurrence times. This algorithm (see appendix B.1) yields proper samples because of the memoryless property of the Exponential. Censoring times were generated from an independent Unif$(0, a)$, where a was choosen to yield a censoring rate of approximatively 20%. The m.l.e.'s for each of the three scenarios were computed from the score functions given in section 5.3.1. This was then repeated 1,000 times and the Monte-Carlo variance estimates of $\hat{\lambda}_j$ ($j = 1, 2, 3$) were computed using the empirical variance formula

$$\widehat{\text{Var}}(\hat{\lambda}_j) = \frac{1}{Q-1}\sum_{q=1}^{Q}\left(\hat{\lambda}_j^q - \bar{\hat{\lambda}}_j\right)^2 , \tag{5.26}$$

where

$$\bar{\hat{\lambda}}_j = \frac{1}{Q}\sum_{q=1}^{Q}\hat{\lambda}_j^q \tag{5.27}$$

and $\hat{\lambda}_j^q$ is the m.l.e. for the j^{th} scenario of the q^{th} iteration and $Q = 1,000$. Finally, $\bar{\hat{\mathcal{I}}}(\hat{\lambda}_j)$'s (for $j = 1, 2, 3$) are computed using the relation $\text{Var}(\hat{\lambda}) = \mathcal{I}^{-1}(\lambda)$. Results are shown in figures 5.3 (for $n = 10$) and 5.4 (for $n = 100$). Both figures are in agreement with the analytical results given at the beginning of section 5.3.2. As differences between the curves of the $\bar{\hat{\mathcal{I}}}(\hat{\lambda}_j)$'s are easier to see, they were plotted instead of the $\widehat{\text{Var}}(\hat{\lambda}_j)$'s in figures 5.3 and 5.4.

Results (5.24) and (5.25), as well as the results of the simulation study, are not surprising and are in agreement with the memoryless property of the Exponential. Since

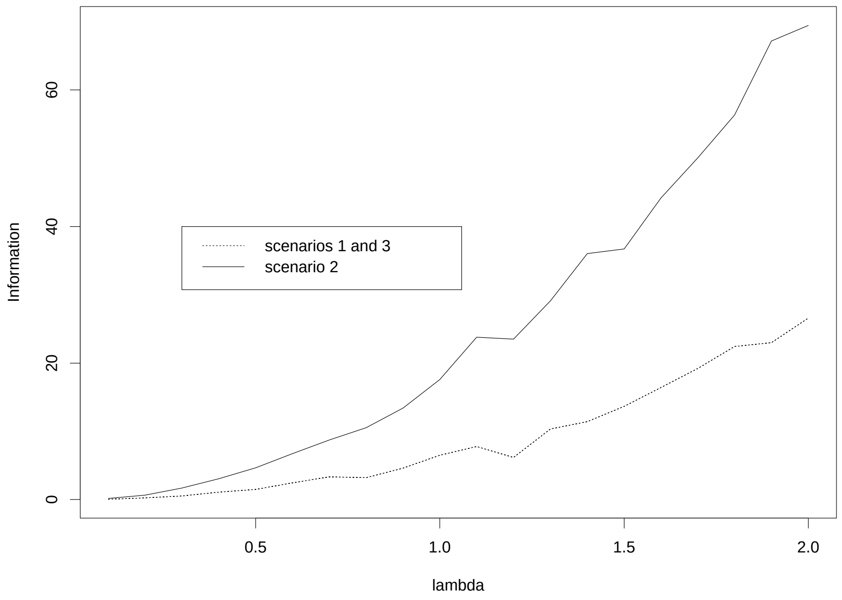

Figure 5.3: Observed information functions for the three scenarios of section 5.2 under the Exponential d.f. with $n = 10$ and ≈20% censoring.

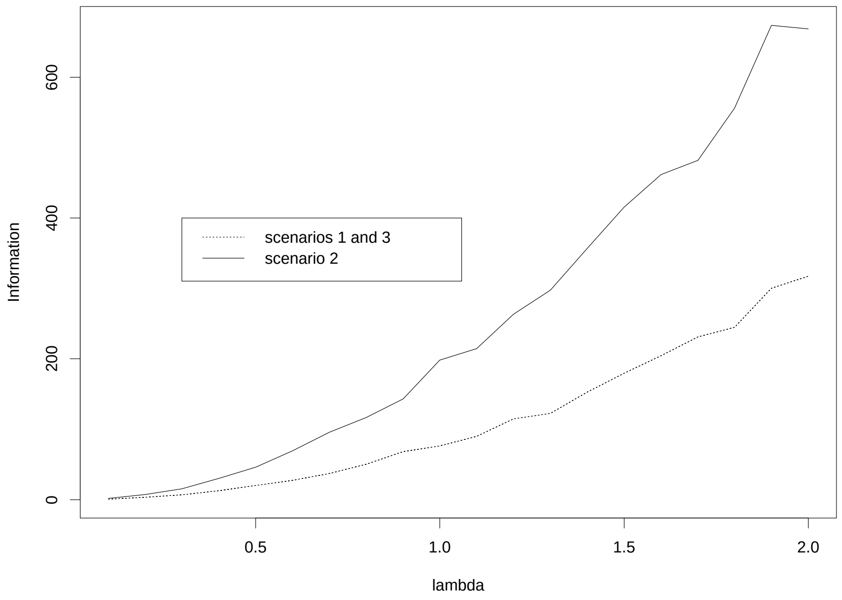

Figure 5.4: Observed information functions for the three scenarios of section 5.2 under the Exponential d.f. with $n = 100$ and ≈20% censoring.

$f(t) = f_{\widetilde{T}}(t)$, intuition suggests that $I(\lambda_1) = I(\lambda_3)$. Similarly, for scenario 2, which involves both $f(t)$ and $f_B(t)$ and where $f(t) = f_B(t)$, intuition indicates that $I(\lambda_2) > I(\lambda_1)$.

5.4 Weibull

The Weibull is perhaps the most widely used distribution in lifetime data analysis. It is more flexible than the Exponential, allowing it to be a better and a more realistic approximation of many lifetime data. Unfortunately, computations required by the scenarios of section 5.2 are considerably more challenging for the Weibull than for the Exponential and it is impossible to analytically derive results when observations are subject to censoring. Therefore, section 5.4.1 looks at the case where the $\widetilde{T}_i$'s are fully observed and where it is possible to analytically compare the information matrices of the three scenarios. Building on the no-censoring comparisons, section 5.4.2 looks at the more realistic case where the $\widetilde{T}_i$'s are subject to censoring and comparisons of the scenarios of section 5.2 are made through simulation studies.

Let $T_1, \ldots, T_n$ be i.i.d. Weibull(λ, β); that is, follow a Weibull distribution with scale parameter λ and shape parameter β. The Weibull density and survivor functions are, respectively,

$$f(t) = \lambda\beta(\lambda t)^{\beta-1} \exp\{-(\lambda t)^{\beta}\} \quad \text{for } t \geq 0 \tag{5.28}$$

and

$$S(t) = \exp\{-(\lambda t)^{\beta}\} \quad \text{for } t \geq 0 \ , \tag{5.29}$$

where $\lambda > 0$ and $\beta > 0$. The k^{th} moment of the Weibull distribution is,

$$E\left(T_i^k\right) = \frac{\Gamma(1 + k/\beta)}{\lambda^k} \ , \tag{5.30}$$

where

$$\Gamma(x) = \int_0^{\infty} u^{x-1} \exp\{-u\} \, du \quad \text{for } x \geq 0 \tag{5.31}$$

is the gamma function (see appendix B.3 for some of the properties of the gamma function).

When $T_i \sim$ Weibull(λ, β), the backward (i.e., B_i) and forward (i.e., $\widetilde{T}_i$) recurrence distributions given by (5.4) become

$$f_B(t) = f_{\widetilde{T}}(t) = \frac{\lambda}{\Gamma(1+1/\beta)} \exp\{-(\lambda t)^\beta\} \quad \text{for } t \geq 0 \,. \tag{5.32}$$

Similarly, when $T_i \sim$ Weibull(λ, β), the length-biased distribution give by (5.3) becomes

$$f_{LB}(t) = \frac{\lambda\beta}{\Gamma(1+1/\beta)}\,(\lambda t)^\beta \exp\{-(\lambda t)^\beta\} \quad \text{for } t \geq 0 \,. \tag{5.33}$$

5.4.1 Without censoring

In this section, the scenarios of section 5.2 are explored and their Fisher information matrices compared when $T_i \sim$ Weibull(λ, β), for $i = 1, \ldots, n$, and the T_i's are uncensored. First, the likelihood and log-likelihood functions, the score vector, observed and Fisher information matrices are computed for each of the three scenarios. Then, the diagonal elements of the three Fisher information matrices are compared at the end of section 5.4.1.

Conditional or scenario 1

For the Weibull distribution, with p.d.f. given by (5.28), (5.5) becomes,

$$L(\lambda, \beta) = \prod_{i=1}^{n} \frac{\lambda\beta(\lambda t_i)^{\beta-1}\exp\{-(\lambda t_i)^\beta\}}{\exp\{-(\lambda b_i)^\beta\}} \,. \tag{5.34}$$

The corresponding log-likelihood function and score vector are,

$$l(\lambda, \beta) = n\ln\lambda + n\ln\beta + (\beta-1)\sum_{i=1}^{n}\ln(\lambda t_i) - \sum_{i=1}^{n}(\lambda t_i)^\beta + \sum_{i=1}^{n}(\lambda b_i)^\beta \tag{5.35}$$

and

$$U(\lambda, \beta) = \Big(U_1(\lambda, \beta), U_2(\lambda, \beta)\Big)' \,, \tag{5.36}$$

where

$$U_1(\lambda,\beta) = n\beta/\lambda - \beta\lambda^{\beta-1}\sum_{i=1}^{n}\left(t_i^\beta - b_i^\beta\right)$$

and

$$U_2(\lambda,\beta) = n/\beta + \sum_{i=1}^{n}\ln(\lambda t_i) - \sum_{i=1}^{n}\left((\lambda t_i)^\beta\ln(\lambda t_i) - (\lambda b_i)^\beta\ln(\lambda b_i)\right) .$$

The observed information matrix corresponding to scenario (5.5) is

$$\mathcal{I}(\lambda,\beta) = [\mathcal{I}(\lambda,\beta)]_{i,j} , \tag{5.37}$$

for $i,j = 1,2$ and where

$$\begin{aligned}
[\mathcal{I}(\lambda,\beta)]_{1,1} &= n\beta/\lambda^2 + \beta(\beta-1)\lambda^{\beta-2}\sum_{i=1}^{n}\left(t_i^\beta - b_i^\beta\right) , \\
[\mathcal{I}(\lambda,\beta)]_{1,2} &= -n/\lambda + \lambda^{\beta-1}\sum_{i=1}^{n}\left(t_i^\beta - b_i^\beta\right) + \beta\lambda^{\beta-1}\sum_{i=1}^{n}\left(t_i^\beta\ln(\lambda t_i) - b_i^\beta\ln(\lambda b_i)\right) , \\
[\mathcal{I}(\lambda,\beta)]_{2,2} &= n/\beta^2 + \lambda^\beta\sum_{i=1}^{n}\left(t_i^\beta\ln^2(\lambda t_i) - b_i^\beta\ln^2(\lambda b_i)\right) .
\end{aligned}$$

Taking the expectation of (5.37) yields the Fisher information matrix

$$I(\lambda,\beta) = E\big(\mathcal{I}(\lambda,\beta)\big) = [I(\lambda,\beta)]_{i,j} , \tag{5.38}$$

with diagonal elements,

$$[I(\lambda,\beta)]_{1,1} = n(\beta^2-\beta+1)/\lambda^2$$

and

$$[I(\lambda,\beta)]_{2,2} = \frac{n}{\beta^2}(1+\gamma^2-2\gamma+\pi^2/6) - \frac{n}{\beta^3}\big(\psi'(1/\beta) + 2\beta\psi(1/\beta) + \psi^2(1/\beta)\big) ,$$

where $\gamma \approx 0.5772157$ is the Euler-Mascheroni constant; see Weisstein (1999). Computation of $E\big([\mathcal{I}(\lambda,\beta)]_{1,1}\big)$ is fairly straightforward. As it requires computing $E\big((\lambda T_i)^\beta\ln^2(\lambda T_i)\big)$ and $E\big((\lambda B_i)^\beta\ln^2(\lambda B_i)\big)$, $E\big([\mathcal{I}(\lambda,\beta)]_{2,2}\big)$ is harder to derive (see appendix B.2 for computation details).

Backward recurrence time or scenario 2

For the Weibull distribution, with p.d.f. given by (5.28), (5.6) becomes,

$$L(\lambda,\beta)=\prod_{i=1}^{n}\frac{\lambda^2\beta(\lambda t_i)^{\beta-1}}{\Gamma(1+1/\beta)}\ \exp\{-(\lambda t_i)^{\beta}\}\ . \tag{5.39}$$

The corresponding log-likelihood function and score vector are,

$$l(\lambda,\beta)=2n\ln\lambda+n\ln\beta-n\ln\Gamma(1+1/\beta)+(\beta-1)\sum_{i=1}^{n}\ln(\lambda t_i)-\sum_{i=1}^{n}(\lambda t_i)^{\beta} \tag{5.40}$$

and

$$U(\lambda,\beta)=\Big(U_1(\lambda,\beta),U_2(\lambda,\beta)\Big)'\ , \tag{5.41}$$

where

$$U_1(\lambda,\beta)=2n/\lambda+n(\beta-1)/\lambda-\beta\lambda^{\beta-1}\sum_{i=1}^{n}t_i^{\beta}$$

and

$$U_2(\lambda,\beta)=2n/\beta+n\psi(1/\beta)/\beta^2+\sum_{i=1}^{n}\ln(\lambda t_i)-\sum_{i=1}^{n}(\lambda t_i)^{\beta}\ln(\lambda t_i)\ ;$$

and where

$$\psi(x)=\frac{\partial}{\partial x}\Big\{\ln\Gamma(x)\Big\} \tag{5.42}$$

is the digamma function (see appendix B.3 for some of the properties of the digamma function).

The observed information matrix corresponding to scenario (5.6) is

$$\mathcal{I}(\lambda,\beta)=[\mathcal{I}(\lambda,\beta)]_{i,j}\ , \tag{5.43}$$

for $i, j = 1, 2$ and where

$$[\mathcal{I}(\lambda,\beta)]_{1,1} = 2n/\lambda^2 + n(\beta-1)/\lambda^2 + \beta(\beta-1)\lambda^{\beta-2}\sum_{i=1}^{n} t_i^{\beta} \ ,$$

$$[\mathcal{I}(\lambda,\beta)]_{1,2} = -n/\lambda + \lambda^{\beta-1}\sum_{i=1}^{n} t_i^{\beta} + \beta\lambda^{\beta-1}\sum_{i=1}^{n}\Big(t_i^{\beta}\ln(\lambda t_i)\Big) \ ,$$

$$[\mathcal{I}(\lambda,\beta)]_{2,2} = 2n/\beta^2 + 2n\psi(1/\beta)/\beta^3 + n\psi'(1/\beta)/\beta^4 + \lambda^{\beta}\sum_{i=1}^{n}\Big(t_i^{\beta}\ln^2(\lambda t_i)\Big) \ .$$

In addition to the digamma function defined in (5.42), the trigamma function $\psi'(x)$ is defined as,

$$\psi'(x) = \frac{\partial^2}{\partial x^2}\Big\{\ln\Gamma(x)\Big\} \tag{5.44}$$

(see appendix B.3 for some of the properties of the trigamma function).

Taking the expectation of (5.43) yields the Fisher information matrix

$$I(\lambda,\beta) = E\big(\mathcal{I}(\lambda,\beta)\big) = [I(\lambda,\beta)]_{i,j} \ , \tag{5.45}$$

with diagonal elements,

$$[I(\lambda,\beta)]_{1,1} = n(\beta^2+1)/\lambda^2$$

and

$$[I(\lambda,\beta)]_{2,2} = \frac{n}{\beta^2}(2+\gamma^2-2\gamma+\pi^2/6) + \frac{2n}{\beta^3}\psi(1/\beta) + \frac{n}{\beta^4}\psi'(1/\beta) \ ,$$

where γ is the Euler-Mascheroni constant. As for (5.38), $E\big([\mathcal{I}(\lambda,\beta)]_{1,1}\big)$ is easily calculated and $E\big([\mathcal{I}(\lambda,\beta)]_{2,2}\big)$ is computed using the same ideas as the ones given in appendix B.2.

Equilibrium or scenario 3

For the Weibull distribution, with p.d.f. given by (5.28), (5.7) becomes,

$$L(\lambda,\beta) = \prod_{i=1}^{n}\frac{\lambda}{\Gamma(1+1/\beta)}\ \exp\{-(\lambda\tilde{t}_i)^{\beta}\} \ . \tag{5.46}$$

The corresponding log-likelihood function and score vector are,

$$l(\lambda,\beta) = n\ln\lambda - n\ln\Gamma(1+1/\beta) - \sum_{i=1}^{n}(\lambda\tilde{t}_i)^{\beta} \tag{5.47}$$

and

$$U(\lambda,\beta) = \Big(U_1(\lambda,\beta), U_2(\lambda,\beta)\Big)' \,, \tag{5.48}$$

where

$$U_1(\lambda,\beta) = \frac{n}{\lambda} - \beta\lambda^{\beta-1}\sum_{i=1}^{n}\tilde{t}_i^{\beta}$$

and

$$U_2(\lambda,\beta) = n/\beta + n\psi(1/\beta)/\beta^2 - \sum_{i=1}^{n}(\lambda\tilde{t}_i)^{\beta}\ln(\lambda\tilde{t}_i) \,.$$

The observed information matrix corresponding to scenario (5.7) is

$$\mathcal{I}(\lambda,\beta) = [\mathcal{I}(\lambda,\beta)]_{i,j} \,, \tag{5.49}$$

for $i,j = 1,2$ and where

$$[\mathcal{I}(\lambda,\beta)]_{1,1} = n/\lambda^2 + \beta(\beta-1)\lambda^{\beta-2}\sum_{i=1}^{n}\tilde{t}_i^{\beta} \,,$$

$$[\mathcal{I}(\lambda,\beta)]_{1,2} = \lambda^{\beta-1}\sum_{i=1}^{n}\tilde{t}_i^{\beta} + \beta\lambda^{\beta-1}\sum_{i=1}^{n}\Big(\tilde{t}_i^{\beta}\ln(\lambda\tilde{t}_i)\Big) \,,$$

$$[\mathcal{I}(\lambda,\beta)]_{2,2} = n/\beta^2 + 2n\psi(1/\beta)/\beta^3 + n\psi'(1/\beta)/\lambda^4 + \lambda^{\beta}\sum_{i=1}^{n}\Big(\tilde{t}_i^{\beta}\ln^2(\lambda\tilde{t}_i)\Big) \,.$$

Taking the expectation of (5.49) yields the Fisher information matrix

$$I(\lambda,\beta) = E\big(\mathcal{I}(\lambda,\beta)\big) = [I(\lambda,\beta)]_{i,j} \,, \tag{5.50}$$

with diagonal elements,

$$[I(\lambda,\beta)]_{1,1} = n\beta/\lambda^2$$

and

$$[I(\lambda,\beta)]_{2,2} = \frac{n}{\beta^2} + \frac{2n}{\beta^3}(1+\beta)\psi(1/\beta) + \frac{n}{\beta^3}\psi^2(1/\beta) + \frac{n}{\beta^4}(1+\beta)\psi'(1/\beta) \,.$$

Comparisons

The scenarios of section 5.2 are compared in the case of the Weibull distribution where the $\widetilde{T}_i$'s are uncensored. Recall that in this case it is possible to make analytical comparisons between the diagonal elements of the three Fisher information matrices derived in section 5.4.1. Hence, from (5.38), (5.45) and (5.50),

$$[I_3(\lambda,\beta)]_{1,1} = \frac{n}{\lambda^2}\beta \leq [I_1(\lambda,\beta)]_{1,1} = \frac{n}{\lambda^2}(\beta^2-\beta+1) < [I_2(\lambda,\beta)]_{1,1} = \frac{n}{\lambda^2}(\beta^2+1)\ , \quad (5.51)$$

with equality iff. $\beta = 1$ and where $[I_j(\lambda,\beta)]_{1,1}$ is the first diagonal element of the Fisher information matrix for the j^{th} scenario, $j = 1, 2, 3$.

Comparisons between the $[I_j(\lambda,\beta)]_{2,2}$'s (the second diagonal element of the Fisher information matrix for the j^{th} scenario) must be done with the help of numerical methods. Using Maple fsolve function to find the roots between (5.38), (5.45) and (5.50) yielded the following results,

$$\begin{aligned}
&[I_1(\lambda,\beta)]_{2,2} < [I_2(\lambda,\beta)]_{2,2} \leq [I_3(\lambda,\beta)]_{2,2} && \text{if } 0 < \beta \leq 0.5911 \\
&[I_1(\lambda,\beta)]_{2,2} \leq [I_3(\lambda,\beta)]_{2,2} < [I_2(\lambda,\beta)]_{2,2} && \text{if } 0.5911 < \beta \leq 1.4448 \qquad (5.52)\\
&[I_3(\lambda,\beta)]_{2,2} < [I_1(\lambda,\beta)]_{2,2} < [I_2(\lambda,\beta)]_{2,2} && \text{if } \beta > 1.4448\ ,
\end{aligned}$$

with equality iff. $\beta = 0.5911$ or 1.4448.

The inequalities given in (5.51) are not surprising and are in agreement with the results of section 5.3.2 for the Exponential. Unfortunately, it is difficult to give any strong interpretation to the inequalities given in (5.52). Analytical comparisons between the $\text{Var}(\hat{\lambda}_j)$'s or the $\text{Var}(\hat{\beta}_j)$'s were not done because of the complexity of the equations involved; a simulation study will be done instead in section 5.4.2.

5.4.2 With censoring

Conditional or scenario 1

For the Weibull distribution, with p.d.f. given by (5.28), (5.5) becomes,

$$L(\lambda,\beta)=\prod_{i=1}^{n}\frac{(\lambda\beta(\lambda t_i)^{\beta-1})^{\delta_i}\exp\{-(\lambda t_i)^{\beta}\}}{\exp\{-(\lambda b_i)^{\beta}\}}\ . \tag{5.53}$$

The corresponding log-likelihood function and score vector are,

$$\begin{aligned} l(\lambda,\beta) &= d\ln\lambda + d\ln\beta + d(\beta-1)\ln\lambda + \\ &\qquad (\beta-1)\sum_{i=1}^{n}\delta_i\ln t_i - \sum_{i=1}^{n}\Big((\lambda t_i)^{\beta}-(\lambda b_i)^{\beta}\Big) \end{aligned} \tag{5.54}$$

and

$$U(\lambda,\beta)=\Big(U_1(\lambda,\beta),U_2(\lambda,\beta)\Big)'\ , \tag{5.55}$$

where

$$\begin{aligned} U_1(\lambda,\beta) &= d/\lambda + d(\beta-1)/\lambda + \beta\lambda^{\beta-1}\sum_{i=1}^{n}\left(t_i^{\beta}-b_i^{\beta}\right)\ , \\ U_2(\lambda,\beta) &= d/\beta + d\ln\lambda + \sum_{i=1}^{n}\delta_i\ln t_i - \sum_{i=1}^{n}\Big((\lambda t_i)^{\beta}\ln(\lambda t_i)-(\lambda b_i)^{\beta}\ln(\lambda b_i)\Big) \end{aligned}$$

and

$$d=\sum_{i=1}^{n}\delta_i\ .$$

The observed information matrix corresponding to scenario (5.5) is

$$\mathcal{I}(\lambda,\beta)=[\mathcal{I}(\lambda,\beta)]_{i,j}\ , \tag{5.56}$$

for $i,j=1,2$ and where

$$\begin{aligned} [\mathcal{I}(\lambda,\beta)]_{1,1} &= d/\lambda^2 + d(\beta-1)/\lambda^2 + \beta(\beta-1)\lambda^{\beta-2}\sum_{i=1}^{n}\left(t_i^{\beta}-b_i^{\beta}\right)\ , \\ [\mathcal{I}(\lambda,\beta)]_{1,2} &= -d/\lambda + \lambda^{\beta-1}\sum_{i=1}^{n}\left(t_i^{\beta}-b_i^{\beta}\right) + \beta\lambda^{\beta-1}\sum_{i=1}^{n}\left(t_i^{\beta}\ln(\lambda t_i)-b_i^{\beta}\ln(\lambda b_i)\right)\ , \\ [\mathcal{I}(\lambda,\beta)]_{2,2} &= d/\beta^2 + \lambda^{\beta}\sum_{i=1}^{n}\left(t_i^{\beta}\ln^2(\lambda t_i)-b_i^{\beta}\ln^2(\lambda b_i)\right)\ . \end{aligned}$$

Backward recurrence time or scenario 2

For the Weibull distribution, with p.d.f. given by (5.28), (5.6) becomes,

$$L(\lambda,\beta)=\prod_{i=1}^{n}\frac{\lambda}{\Gamma(1+1/\beta)}\left(\lambda\beta(\lambda t_i)^{\beta-1}\right)^{\delta_i}\exp\{-(\lambda t_i)^{\beta}\}\ . \tag{5.57}$$

The corresponding log-likelihood function and score vector are,

$$\begin{aligned} l(\lambda,\beta)=(n+d)\ln\lambda-n\ln\Gamma(1+1/\beta)+d\ln\beta+\\ d(\beta-1)\ln\lambda+(\beta-1)\sum_{i=1}^{n}\delta_i\ln t_i-\sum_{i=1}^{n}(\lambda t_i)^{\beta}\end{aligned} \tag{5.58}$$

and

$$U(\lambda,\beta)=\Big(U_1(\lambda,\beta),U_2(\lambda,\beta)\Big)'\ , \tag{5.59}$$

where

$$U_1(\lambda,\beta)=(n+d\beta)/\lambda-\beta\lambda^{\beta-1}\sum_{i=1}^{n}t_i^{\beta}$$

and

$$U_2(\lambda,\beta)=(n+d)/\beta+d\ln\lambda+n\psi(1/\beta)/\beta^2+\sum_{i=1}^{n}\delta_i\ln t_i-\sum_{i=1}^{n}(\lambda t_i)^{\beta}\ln(\lambda t_i)\ .$$

The observed information matrix corresponding to scenario (5.6) is

$$\mathcal{I}(\lambda,\beta)=[\mathcal{I}(\lambda,\beta)]_{i,j}\ , \tag{5.60}$$

for $i,j=1,2$ and where

$$[\mathcal{I}(\lambda,\beta)]_{1,1}=(n+d\beta)/\lambda^2+\beta(\beta-1)\lambda^{\beta-2}\sum_{i=1}^{n}t_i^{\beta}\ ,$$

$$[\mathcal{I}(\lambda,\beta)]_{1,2}=-d/\lambda+\lambda^{\beta-1}\sum_{i=1}^{n}t_i^{\beta}\Big(\beta\ln(\lambda t_i)+1\Big)\ ,$$

$$[\mathcal{I}(\lambda,\beta)]_{2,2}=\frac{n+d}{\beta^2}+\frac{n}{\beta^3}\left(2\psi(1/\beta)+\frac{1}{\beta}\psi'(1/\beta)\right)+\lambda^{\beta}\sum_{i=1}^{n}\Big(t_i^{\beta}\ln^2(\lambda t_i)\Big)\ .$$

Equilibrium or scenario 3

For the Weibull distribution, with p.d.f. given by (5.28), (5.7) becomes,

$$L(\lambda,\beta)=\prod_{i=1}^{n}\frac{\lambda}{\Gamma(1+1/\beta)}\left(\exp\{-(\lambda\tilde{t}_i)^\beta\}\right)^{\delta_i}\left(\frac{1}{\lambda\beta}\,\Gamma(1/\beta,(\lambda\tilde{t}_i)^\beta)\right)^{1-\delta_i}\ , \tag{5.61}$$

where

$$\Gamma(x,y)=\int_y^\infty u^{x-1}\exp\{-u\}\,du \tag{5.62}$$

is the "upper" incomplete gamma function, or simply the incomplete gamma function (see appendix B.3 for some of the properties of the incomplete gamma function). The corresponding log-likelihood function and score vector are, respectively,

$$\begin{aligned} l(\lambda,\beta)=d\ln\lambda+(d-n)\ln\beta-n\ln\Gamma(1+1/\beta)-\\ \lambda^\beta\sum_{i=1}^{n}\delta_i\tilde{t}_i^\beta+\sum_{i=1}^{n}(1-\delta_i)\ln\Gamma(1/\beta,(\lambda\tilde{t}_i)^\beta)\end{aligned} \tag{5.63}$$

and

$$U(\lambda,\beta)=\Big(U_1(\lambda,\beta),U_2(\lambda,\beta)\Big)'\ , \tag{5.64}$$

where

$$U_1(\lambda,\beta)=\frac{d}{\lambda}-\beta\lambda^{\beta-1}\sum_{i=1}^{n}\delta_i\tilde{t}_i^\beta-\beta\sum_{i=1}^{n}(1-\delta_i)\,\frac{\tilde{t}_i\exp\{-(\lambda\tilde{t}_i)^\beta\}}{\Gamma(1/\beta,(\lambda\tilde{t}_i)^\beta)}$$

and

$$\begin{aligned} U_2(\lambda,\beta)=\frac{d}{\beta}+\frac{n}{\beta^2}\psi(1/\beta)-\sum_{i=1}^{n}\delta_i(\lambda\tilde{t}_i)^\beta\ln(\lambda\tilde{t}_i)-\\ \sum_{i=1}^{n}\frac{(1-\delta_i)}{\Gamma(1/\beta,(\lambda\tilde{t}_i)^\beta)}\left(\frac{\Gamma'(1/\beta,(\lambda\tilde{t}_i)^\beta)}{\beta^2}+\lambda\tilde{t}_i\exp\{-(\lambda\tilde{t}_i)^\beta\}\ln(\lambda\tilde{t}_i)\right)\ ,\end{aligned}$$

where

$$\Gamma'(x,y)=\frac{\partial}{\partial x}\Big\{\Gamma(x,y)\Big\}=\int_y^\infty u^{x-1}\,\ln u\,\exp\{-u\}\,du\ .$$

The observed information matrix corresponding to scenario (5.7) is

$$\mathcal{I}(\lambda,\beta) = [\mathcal{I}(\lambda,\beta)]_{i,j}\ , \tag{5.65}$$

for $i,j = 1,2$ and where

$$\begin{aligned}
[\mathcal{I}(\lambda,\beta)]_{1,1} = {} & \frac{d}{\lambda^2} + \beta(\beta-1)\lambda^{\beta-2}\sum_{i=1}^{n}\delta_i\tilde{t}_i^{\beta} - \beta^2\lambda^{\beta-1}\sum_{i=1}^{n}(1-\delta_i)\,\frac{\tilde{t}_i^{\beta+1}\exp\{-(\lambda\tilde{t}_i)^{\beta}\}}{\Gamma(1/\beta,(\lambda\tilde{t}_i)^{\beta})} + \\
& \beta^2\sum_{i=1}^{n}(1-\delta_i)\left(\frac{\tilde{t}_i\exp\{-(\lambda\tilde{t}_i)^{\beta}\}}{\Gamma(1/\beta,(\lambda\tilde{t}_i)^{\beta})}\right)^2\ , \\
[\mathcal{I}(\lambda,\beta)]_{1,2} = {} & \lambda^{\beta-1}\sum_{i=1}^{n}\delta_i\tilde{t}_i^{\beta}\Big(1+\beta\ln(\lambda\tilde{t}_i)\Big) + \\
& \sum_{i=1}^{n}(1-\delta_i)\,\frac{\tilde{t}_i\exp\{-(\lambda\tilde{t}_i)^{\beta}\}}{\Gamma(1/\beta,(\lambda\tilde{t}_i)^{\beta})}\Big(1-\beta(\lambda\tilde{t}_i)^{\beta}\ln(\lambda\tilde{t}_i)\Big) + \\
& \beta\sum_{i=1}^{n}(1-\delta_i)\,\frac{\tilde{t}_i\exp\{-(\lambda\tilde{t}_i)^{\beta}\}}{\Gamma^2(1/\beta,(\lambda\tilde{t}_i)^{\beta})}\left(\frac{\Gamma'(1/\beta,(\lambda\tilde{t}_i)^{\beta})}{\beta^2} + \lambda\tilde{t}_i\ln(\lambda\tilde{t}_i)\exp\{-(\lambda\tilde{t}_i)^{\beta}\}\right)\ ,
\end{aligned}$$

$$\begin{aligned}
[\mathcal{I}(\lambda,\beta)]_{2,2} = {} & \frac{d}{\beta^2} + \frac{2n}{\beta^3}\psi(1/\beta) + \frac{n}{\beta^4}\psi'(1/\beta) + \sum_{i=1}^{n}\delta_i(\lambda\tilde{t}_i)^{\beta}\ln^2(\lambda\tilde{t}_i) - \\
& \frac{1}{\beta^4}\sum_{i=1}^{n}\frac{1-\delta_i}{\Gamma(1/\beta,(\lambda\tilde{t}_i)^{\beta})}\Big(\Gamma''(1/\beta,(\lambda\tilde{t}_i)^{\beta}) + 2\beta\Gamma'(1/\beta,(\lambda\tilde{t}_i)^{\beta}) + \\
& \quad \beta^3(\lambda\tilde{t}_i)\ln^2(\lambda\tilde{t}_i)\exp\{-(\lambda\tilde{t}_i)^{\beta}\}\big(1+\beta(\lambda\tilde{t}_i)^{\beta}\big)\Big) - \\
& \frac{1}{\beta^4}\sum_{i=1}^{n}\frac{1-\delta_i}{\big(\Gamma(1/\beta,(\lambda\tilde{t}_i)^{\beta})\big)^2}\Big(\Gamma'(1/\beta,(\lambda\tilde{t}_i)^{\beta}) + \beta^2\lambda\tilde{t}_i\ln(\lambda\tilde{t}_i)\exp\{-(\lambda\tilde{t}_i)^{\beta}\}\Big)^2\ ,
\end{aligned}$$

where

$$\Gamma''(x,y) = \frac{\partial^2}{\partial x^2}\Big\{\Gamma(x,y)\Big\} = \int_y^{\infty} u^{x-1}(\ln u)^2\exp\{-u\}\,du\ .$$

Note that the computation of the elements of (5.65) makes use of numerous properties of the family of gamma functions (see appendix B.3).

5.4.3 Simulation study

The complexity of the formulas in section 5.4.2 makes it impossible to analytically derive comparisons, as was done in sections 5.3.2 (Exponential distribution) and 5.4.1 (Weibull distribution without censoring). Therefore, a simulation study was done instead to compare the three scenarios of section 5.4.2. The first part of this section describes how the simulation study was implemented, as well as the challenges and difficulties encountered in computing the m.l.e. for the equilibrium scenario. The second part of this section describes the results of the simulation study.

Description of simulation study

The first element of the simulation study is to generate the random variables (b_i, t_i, δ_i) for $i = 1, \ldots, n$, where $n = 100$ was chosen for all simulations. This was done using algorithm 5.1.

Algorithm 5.1 (generating random observations):

Step 1: For the given values of (λ, β), generate $t_1, \ldots, t_{100}$ i.i.d. $f_{LB}(t)$, where $f_{LB}(t)$ was given by (5.33); To facilitate computer simulations, note that if $T \sim f_{LB}(t)$ then $Y = T^\beta \sim \text{Gamma}(1/\beta + 1, \lambda^\beta)$;

Step 2: For $i = 1, \ldots, 100$, generate $b_i \sim \text{Unif}(0, t_i)$, where $B \mid T = t \sim \text{Unif}(0, t)$ is the conditional distribution of B, given t;

Step 3: Compute $\tilde{t}_i = t_i - b_i$ for $i = 1, \ldots, 100$;

Step 4: Apply censoring to $\tilde{t}_1, \ldots, \tilde{t}_{100}$; This was done in a much simpler way than in the algorithm given in section B.1 (e.g., see the justification of the choice of a in that algorithm); If $\tilde{t}_i > q_{1-\alpha}$, where $q_{1-\alpha}$ is the $(1-\alpha)$ empirical quantile of $\tilde{t}_1, \ldots, \tilde{t}_{100}$, then $\delta_i = 0$ and $\tilde{t}_i = q_{1-\alpha}$, otherwise $\delta_i = 1$; In other words, we used type 2 censoring; For the simulations of section 5.4.3, $\alpha = 0.2$ was chosen;

Step 5: Compute the new values of $t_i = \tilde{t}_i + b_i$ for $i = 1, \ldots, 100$, where $\tilde{t}_i$ was redefined in step 4; Note that these t_i's account for censoring.

The m.l.e. for the conditional scenario of section 5.4.2 was computed using Newton-Raphson algorithm, with score and observed information matrix given by (5.55) and (5.56). Note that the Newton-Raphson method, which uses derivatives, was originally developed to find the roots of an equation or a system of equations. Press et al. (2002) give additional explanations on the Newton-Raphson algorithm and how it might fail to converge. Regardless of its origin, this algorithm is frequently used to do multidimensional function maximization (or minimization), as was done here. The stopping criterion used in section 5.4.3 was inspired by the C++ function NR::mnewt from Press et al. (2002), section 9.6. Hence, convergence was declared if $\| \mathcal{I}^{-1}(\mu, \sigma)\, U(\mu, \sigma) \| <$ tolerance, where tolerance was fixed at the start of the simulation study, and where the re-parameterization $\mu = -\log \lambda$ and $\sigma = \log \beta$ was preferred as it simplifies the maximization procedure since there is no constraint. The other element to consider when using Newton-Raphson, is the choice of a starting point for the algorithm. An option with the simulations of section 5.4.3, would be to use standard statistical software (e.g., the S-Plus survreg function) to fit a model using only the values of (t_i, δ_i)'s; see algorithm 5.1. This would obviously yield biased estimates of λ and β as the T_i's follow a length-biased Weibull and not a Weibull distribution. Nevertheless, these could be used as a starting point for the algorithm. However, to speed up the simulations, this option was not retained and the values of (λ, β) used in algorithm 5.1 were given as the starting point for the Newton-Raphson algorithm. In rare cases, the Newton-Raphson algorithm with this starting point failed to converge to a global maximum. In such cases, a grid of (λ, β) points was generated and the log-likelihood function, as given by (5.54), was computed at each of these points. The Newton-Raphson algorithm was then re-started at the point that yielded the largest value for the log-likelihood function. Using this strategy, the Newton-Raphson algorithm converged in more than 95% of simulations. It was not re-started with another starting point if it failed again, and that particular simulation was dismissed.

The m.l.e. for the backward recurrence time scenario of section 5.4.2 was also computed using Newton-Raphson algorithm. Obviously, the score and observed information matrix given by (5.55) and (5.56) were replaced by (5.59) and (5.60), respectively. The same criterion discussed in the previous paragraph to declare convergence was also used here. A slightly different parameterization for (μ, σ) was chosen; that is, $\mu = -\log\lambda$ and $\sigma = \beta$. In addition, the discussion regarding the choice of a starting point for the Newton-Raphson algorithm and what to do if the algorithm fails to converge to a global maximum applies here as well. Using the same strategy as the one used for computing the the m.l.e. of the conditional scenario, yielded a convergence rate of approximately 99% for this second scenario.

Computation of the m.l.e. for the equilibrium scenario of section 5.4.2 was much more complicated and challenging that for the two previous scenarios. The log-likelihood function, given by (5.63), had one or more singularities for many datasets. This implied that Newton-Raphson would not converge in these cases. Hence, a purely numerical maximization procedure was preferred. In addition, the re-parameterization $\mu = -\log\lambda$ with $\sigma = \log\beta$ was chosen as it allows for maximization without constraint. For simplicity and ease of use, the S-Plus nlminb function was preferred in the present simulation study. The nlminb function does not need derivatives as it is based on quasi-Newton methods, also called variable metric methods; see Press et al. (2002), section 10.7. However, other numerical maximization methods could have been used instead; see section 6.2 or Press et al. (2002), chapter 10. As the values of $t_1, \ldots, t_n$ would not be known if this was not a simulation study, it is difficult to justify using the survreg function to obtain a starting point for the maximization algorithm. As for the conditional and backward scenarios, the values of (λ, β) used in algorithm 5.1 were given as the starting point. If nlminb failed to converge to a global maximum with this starting point, a grid of (λ, β) points was generated. The log-likelihood function, as given by (5.63), was computed at each of these points and the nlminb function re-started at the point which yielded the largest value of the log-likelihood function. Thus, the same strategy, as the one described in the previous two paragraphs, was used if the algorithm failed to converge to a global maximum. The difference is that the S-plus nlminb function had to be re-started much more frequently and failed to converge at all in many more cases than for the two other scenarios of section 5.4.2. For $1 < \beta < 5$, this was not a major problem as approximately 90% of the simulations (including those that had been re-started) converged to a global maximum. However, for other values of β this rate went down to less than 60% in some cases (e.g., $(\lambda, \beta) = (15, 1)$ and $(\lambda, \beta) = (20, 5.5)$).

Algorithm 5.2 (simulation study):

Step 1: For (λ, β), $n = 100$ and 20% censoring, generate the dataset according to algorithm 5.1;

Step 2: Compute the m.l.e.'s for the three scenarios of section 5.4.2 using the techniques described in the three previous paragraphs;

Step 3: Compute $\mathcal{I}(\hat{\lambda}, \hat{\beta})$ for all three scenarios (using (5.56), (5.60) and (5.65));

Step 4: Repeat steps 1 through 3 a 1,000 times[1] and compute $\bar{\hat{\lambda}}$, $\widehat{\overline{\mathrm{Var}}}(\hat{\lambda})$ and $\hat{\bar{\mathcal{I}}}(\hat{\lambda})$ using (5.26) and (5.27); Similarly, compute $\bar{\hat{\beta}}$, $\widehat{\overline{\mathrm{Var}}}(\hat{\beta})$ and $\hat{\bar{\mathcal{I}}}(\hat{\beta})$;

Step 5: Repeat steps 1 through 4 for all pairwise combinations of $\lambda = 5$, 10, 15, 20 and 25 and $\beta = 1$, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5 and 5.5.

Results of simulation study

The results of the simulation study are shown in figures 5.5–5.9.

Figure 5.5 shows averages of $(\hat{\lambda}, \hat{\beta})$ for the 50 simulations of section 5.4.3. These averages were computed using (5.27), where $Q = 1,000$; see step 4 of algorithm 5.2. The top first quarter of figure 5.5 contains the true values of (λ, β) used in the 50 simulations. The botton half of figure 5.5 indicates that both the conditional and backward recurrence

[1] If a simulation failed to converge it was replace by another one, so the total number of simulations would remain 1,000.

time scenarios yield unbiased estimates of (λ, β). However, this is not the case of the equilibrium scenario; see second top quarter of figure 5.5. More precisely, the estimates of λ have little to no bias, but the bias of $\hat{\beta}$ increases with increasing values of β, regardless of the value of λ. It is difficult not to make a connection between the bias and the convergence problems mentioned previously. However, further research would be needed to formally establish that connection and to gain a better understanding on the cause of this bias.

Figures 5.6 and 5.7 show the empirical variances of $(\hat{\lambda}, \hat{\beta})$ for the 50 simulations of section 5.4.3. From figure 5.6, $\widehat{\widehat{\mathrm{Var}}}(\hat{\beta})$ for the equilibrium scenario is much greater than for the conditional and backward recurrence time scenarios. Although it is difficult to see from the graphics, $\widehat{\widehat{\mathrm{Var}}}(\hat{\beta})$ for the backward recurrence time scenario is smaller than for the conditional scenario for all of the 50 simulations. From figure 5.7, similar comments to those made regarding the values of $\widehat{\widehat{\mathrm{Var}}}(\hat{\beta})$ also apply to the different values of $\widehat{\widehat{\mathrm{Var}}}(\hat{\lambda})$. Hence, $\widehat{\widehat{\mathrm{Var}}}(\hat{\lambda})$ for the equilibrium scenario is greater, and for many simulations much greater, than for the conditional and backward recurrence time scenarios. Looking at figure 5.7, it is easier to see that $\widehat{\widehat{\mathrm{Var}}}(\hat{\lambda})$ for the backward recurrence time scenario is smaller than for the conditional scenario for all of the 50 simulations.

Figures 5.8 and 5.9 show the average values of $\mathcal{I}(\hat{\lambda}, \hat{\beta})]_{i,i}$ $(i = 1, 2)$ for the 50 simulations of section 5.4.3. These averages are based on the 1,000 repetitions mentioned in step 4 of algorithm 5.2. Let j designate the j^{th} scenario of section 5.4.2 $(j = 1, 2$ or $3)$. Hence, $j = 1$ — conditional or scenario 1; $j = 2$ — backward recurrence time or scenario 2; and $j = 3$ — equilibrium or scenario 3. Then

$$[\mathcal{I}_3(\hat{\lambda}, \hat{\beta})]_{i,i} < [\mathcal{I}_1(\hat{\lambda}, \hat{\beta})]_{i,i} < [\mathcal{I}_2(\hat{\lambda}, \hat{\beta})]_{i,i} \tag{5.66}$$

for $i = 1, 2$ and all but one of the 50 simulations. The only exception is for the $(\lambda = 25, \beta = 1)$ combination, where $[\mathcal{I}_1(\hat{\lambda}, \hat{\beta})]_{1,1} < [\mathcal{I}_3(\hat{\lambda}, \hat{\beta})]_{1,1} < [\mathcal{I}_2(\hat{\lambda}, \hat{\beta})]_{1,1}$. Looking at figures 5.6 and 5.7, these results are not surprising since $\widehat{\mathrm{Var}}(\hat{\lambda}, \hat{\beta}) = \mathcal{I}^{-1}(\hat{\lambda}, \hat{\beta})$. These results are in agreement with the analytical comparisons given by (5.51), but not with those given by (5.52); see section 5.4.1. However, it is impossible to tell if the discrepancy between (5.66) and (5.52) is due to the presence of censored observations or to the converge problems of the equilibrium scenario. As mentioned previously, further research on the cause and possible solutions of the converge problems of the equilibrium scenario is needed.

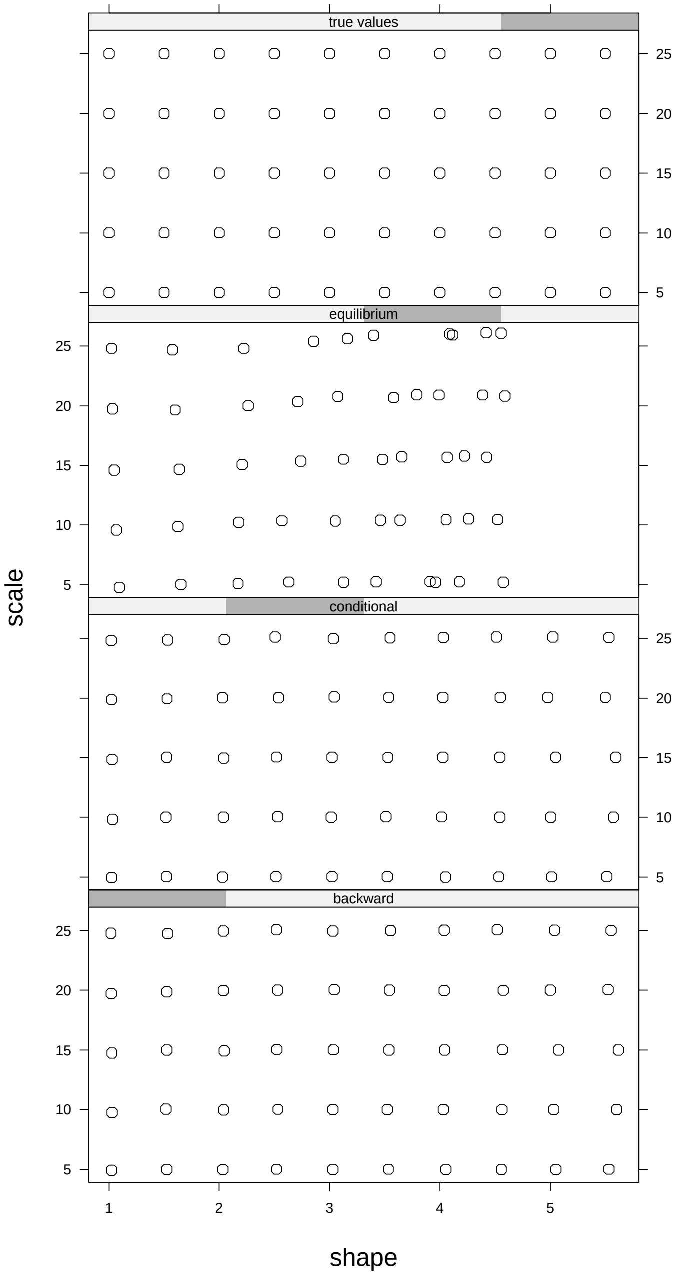

Figure 5.5: Values of $(\bar{\hat{\lambda}}, \bar{\hat{\beta}})$ for the 50 simulations of section 5.4.3.

The true values of (λ, β) used in the simulations are shown at the top for comparisons.

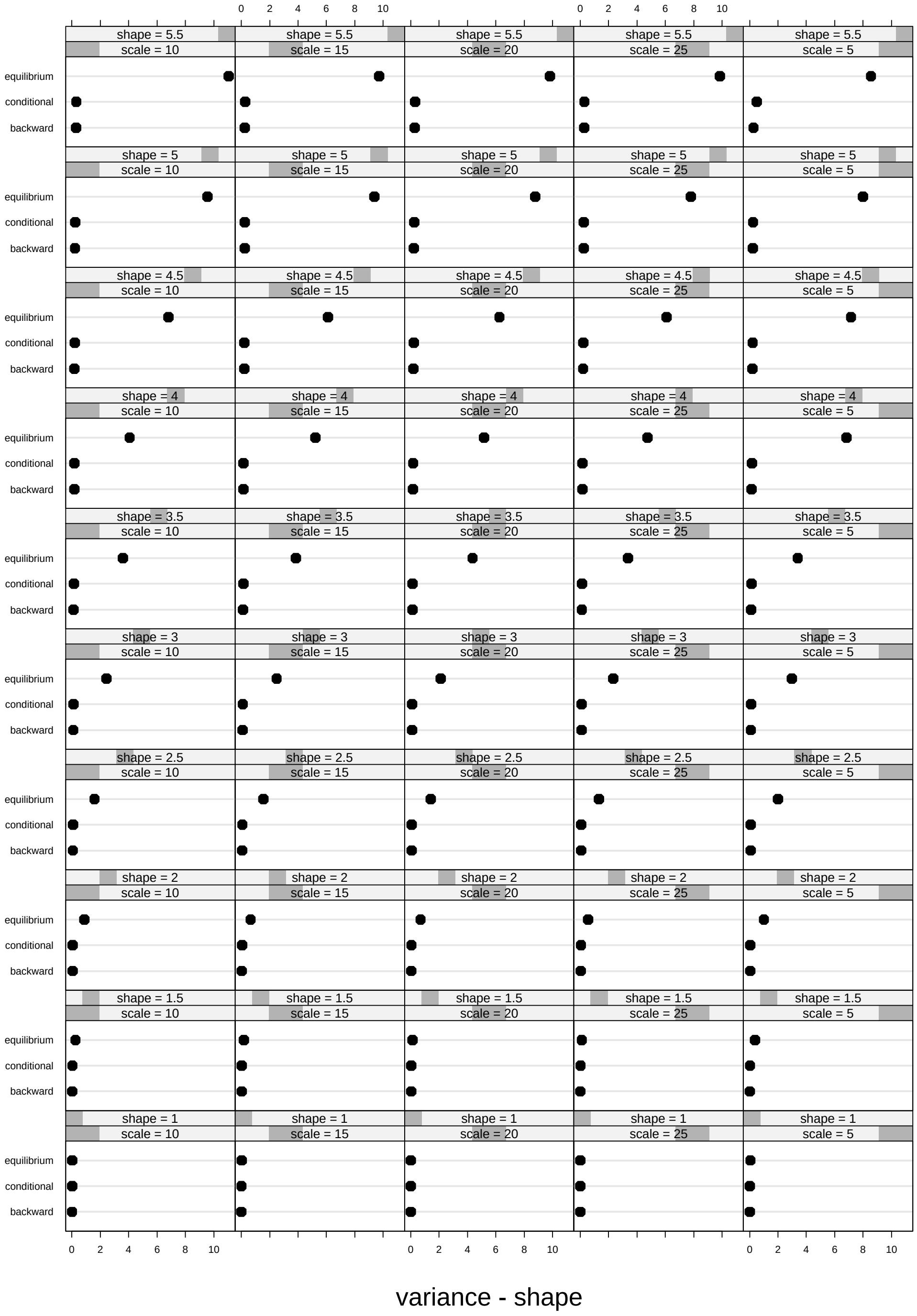

Figure 5.6: Values of $\widehat{\mathrm{Var}}(\hat{\beta})$ for the 50 simulations of section 5.4.3.

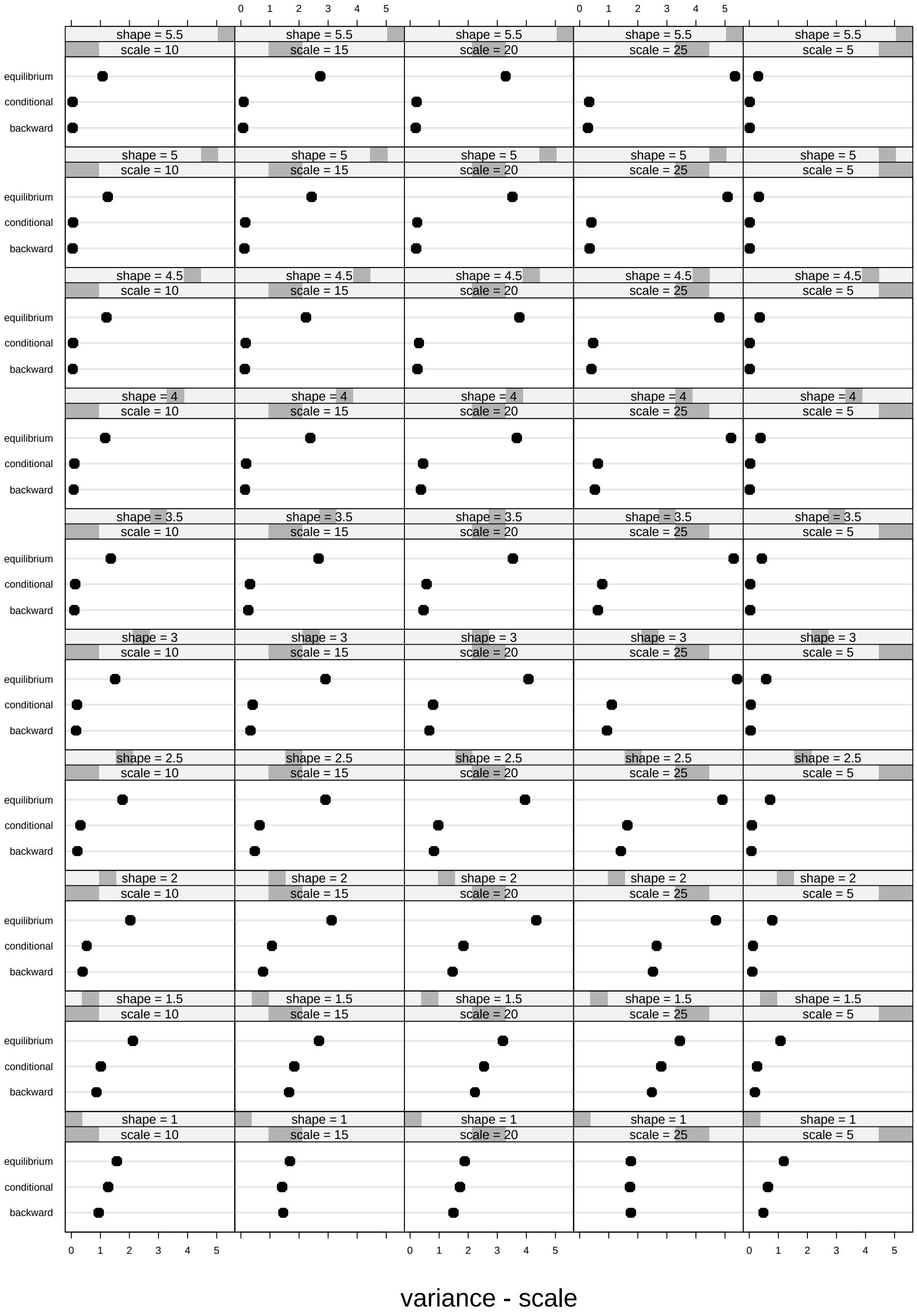

Figure 5.7: Values of $\widehat{\mathrm{Var}}(\hat{\lambda})$ for the 50 simulations of section 5.4.3.

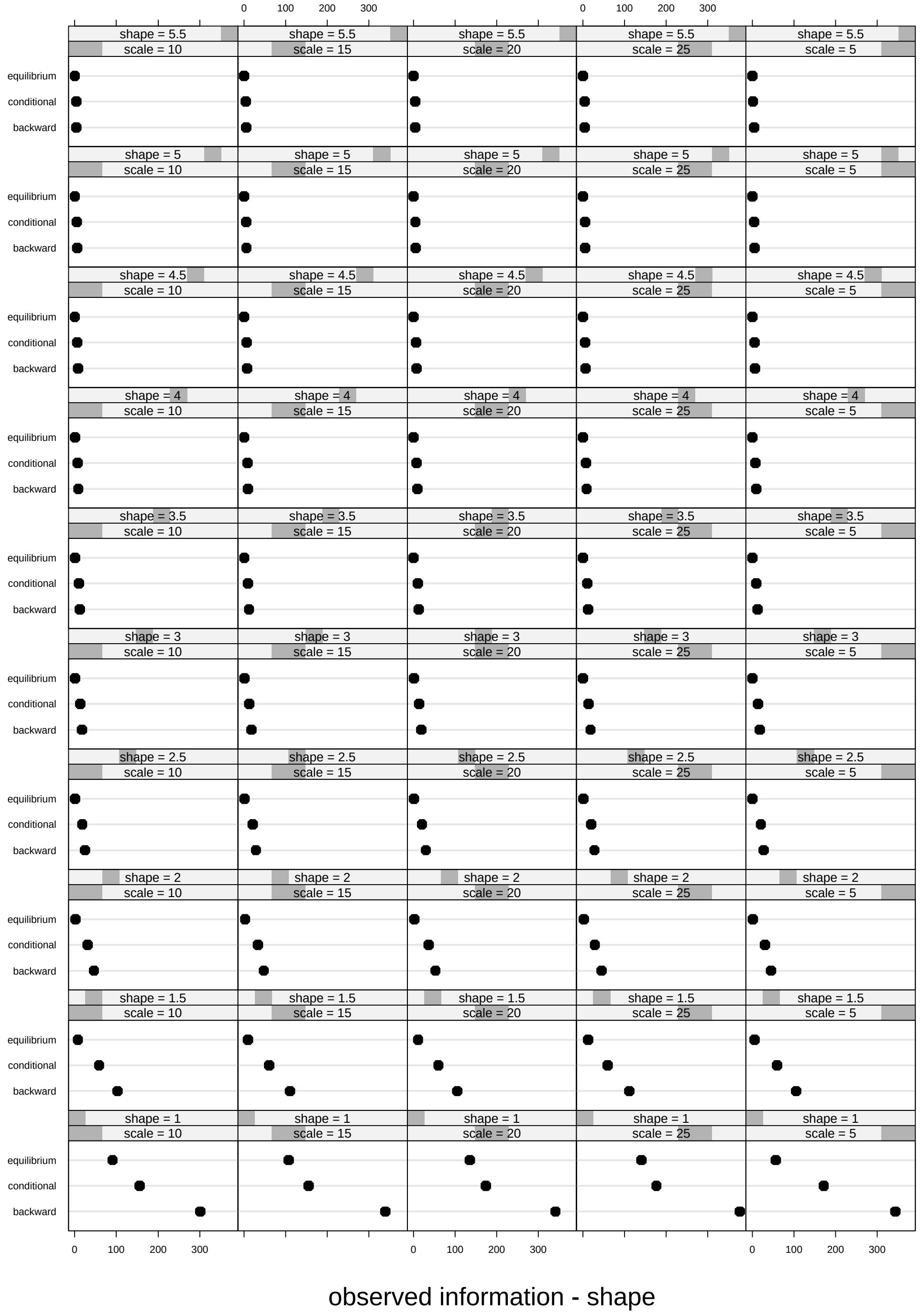

Figure 5.8: Values of $[\hat{\bar{\mathcal{I}}}(\hat{\lambda}, \hat{\beta})]_{2,2}$ for the 50 simulations of section 5.4.3.

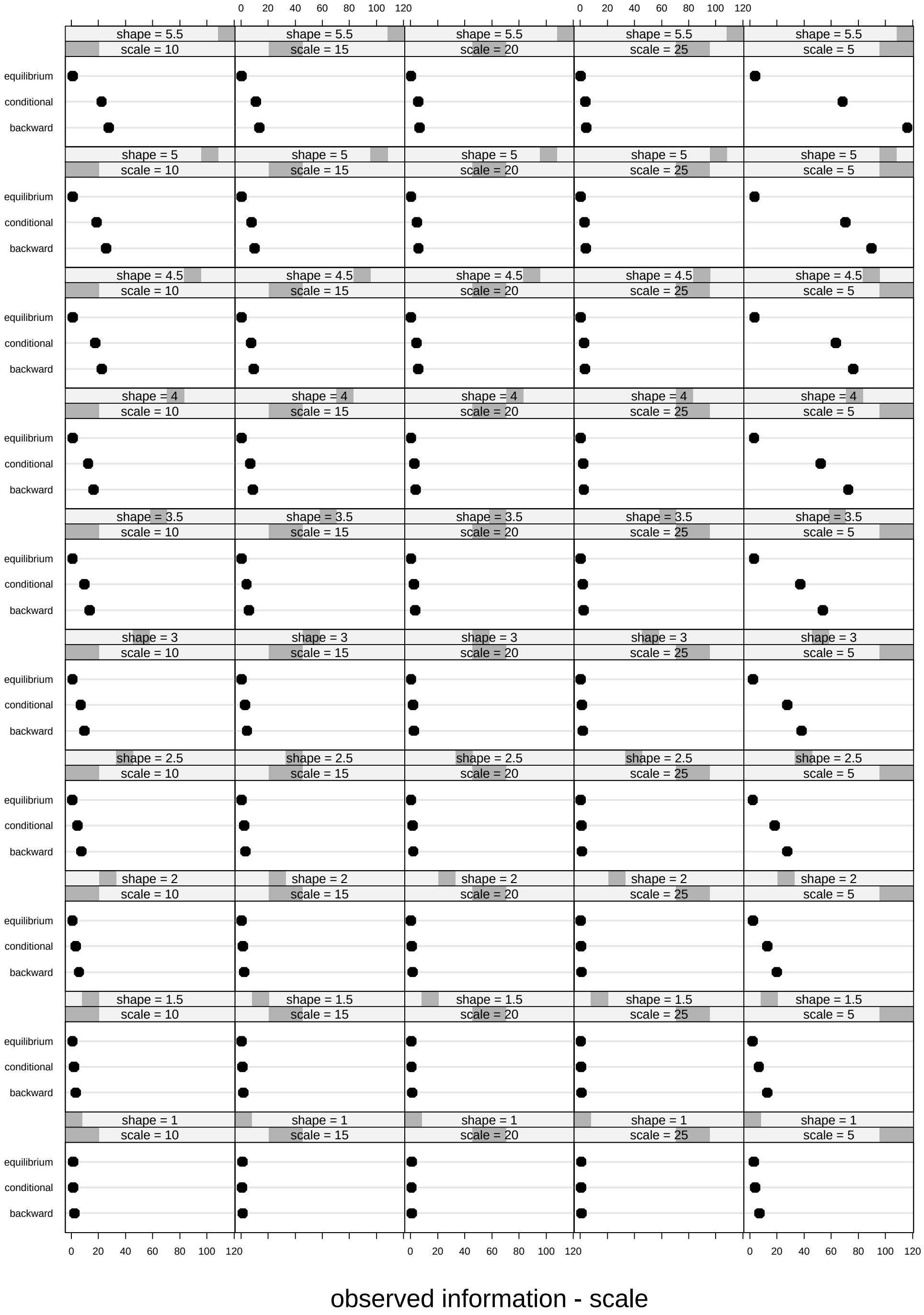

Figure 5.9: Values of $[\hat{\bar{\mathcal{I}}}(\hat{\lambda}, \hat{\beta})]_{1,1}$ for the 50 simulations of section 5.4.3.

Chapter 6

Conclusion and future research

6.1 Summary of scientific contributions

As stated at the beginning of chapter 1, the main objectives of our research have been:

1. To develop statistical methods for carrying out analytical inference from longitudinal survey data when the sampling design is uninformative. These methods account for the correlation between observations and, more generally, for the use of complex survey designs;

2. To propose and compare different parametric statistical models that allow for the inclusion of left-truncated observations; To discuss the conditional partial-likelihood, a semi-parametric method which can handle left-truncated sojourns.

Secondary objectives have been:

3. To shed some light on the controversial topic of the use and the role of sampling weights when carrying out analytical inference from longitudinal survey data. This was done using the notions of uninformative and ignorable sampling designs;

4. To discuss, exemplify and summarize the particularities or unique feature of longitudinal surveys, and to describe their impact on statistical inference.

The most important novelties in this thesis are with respect to item 1. In chapter 3 and, particularly in section 3.2, we were successful in extending the methods introduced by Lee et al. (1992), Wei et al. (1989), Liang et al. (1993), Spiekerman & Lin (1998) and Lin et al. (2000) to the context of analytical inference from longitudinal survey data. This requires allowing for both clusters of various sizes and different baseline hazard functions. Proposed robust estimators for $\mathrm{Var}(\hat{\beta})$ and $\mathrm{Var}(\hat{\Lambda}_{0h}(t,\hat{\beta}))$ are, respectively, given by (3.19) and (3.38).

In addition, weighted versions of the robust estimators for $\mathrm{Var}(\hat{\beta})$ and $\mathrm{Var}(\hat{\Lambda}_{0h}(t,\hat{\beta}))$ were derived in section 3.4.2. These weighted robust variance estimators, given by (3.75) and (3.77), yield similar results to the ones proposed by Binder (1992) and, in particular, by Lin (2000), when applied to the SIPP and SLID data of chapter 4.

In chapter 4, the methods proposed in chapter 3 were applied to analyses of spells on the food stamps program (see section 4.1.1) and of spells without health insurance (see section 4.1.2) using the SIPP data. They were also applied to the analysis of jobless spells (see section 4.2.1) using the SLID data. In all these analyses the methods proposed in chapter 3 performed extremely well.

With respect to item 2 of the previous page, three models or scenarios that allow for the inclusion of left-truncated observations were proposed in section 5.2; they are: the conditional, backward recurrence time and equilibrium scenarios. These scenarios were studied and compared for the Exponential distribution in section 5.3, and analytical comparisons were derived in section 5.3.2. In summary, under the Exponential distribution assumption, the conditional and equilibrium scenarios yield the same results with respect to $I(\lambda)$ and $\mathrm{Var}(\hat{\lambda})$; however, both of these yield smaller values of $I(\lambda)$ and larger values of $\mathrm{Var}(\hat{\lambda})$ than the backward recurrence time scenario for all values of λ, sample sizes and number of censored observations. In section 5.4, the same three scenarios were studied and compared under the Weibull distribution. First, this was done when all observations are uncensored. The results of the analytical comparisons of $[\mathcal{I}(\lambda,\beta)]_{1,1}$ and $[\mathcal{I}(\lambda,\beta)]_{2,2}$ were given by (5.51) and (5.52); see section 5.4.1. Section 5.4.2 allows for both left-truncated and right-censored observations and simulations studies were done in section 5.4.3 to compare the three scenarios.

Item 3 of page 173 was the topic of section 3.3 and, in particular, section 3.3.2. The notions of uninformative and ignorable sampling designs were first introduced. Then, the pros and cons of weighted and unweighted analyses were discussed. As mentioned in section 3.3.2, the point of view taken in this thesis is to use unweighted analyses when the goal is analytical inference for parameters of a super-population or a probabilistic model and when the sampling design is uninformative or ignorable; otherwise, use weighted analyses. This is in agreement with the recommendations of Pfeffermann & Smith (1985).

Finally, with respect to item 4 of page 173, the particularities of longitudinal surveys and their implications were discussed and exemplified in chapter 2. These particularities, which are also the topic of Binder (1998) and of many articles contained in the book edited by Kasprzyk et al. (1989), include: the seam-effect, the prevalence of left-truncated observations, attrition or loss to follow-up as time goes on, the dynamic nature of the population under study, the difficulties of defining "proper" sampling weights and the importance of analytical inference in longitudinal surveys. Some of the issues raised concerning these particularities, in particular the ones of sections 2.4, 2.6 and 2.7, played a key role in the discussion given in section 3.3.2.

6.2 Areas for future research

The following paragraphs give some areas where future research is needed.

In sections 4.1.1, 4.1.2 and 4.2.1, diagnostic model checking was performed under the independence assumption. This was not a major problem as "robust" and "naive" variance estimators were close to one another. However, finding diagnostic tools that can account for correlated observations would be very useful. This could be done by adjusting existing methods to account for correlation, as it was done for the methods of chapter 3.

Weighted versions of the proposed variance estimators of section 3.2 were derived in section 3.4.2. However, further research is needed to study their asymptotic properties. In particular, to what exact quantities do they converge under both finite and super-population models? This would help in the comparisons of these weighted variance estimators of section 3.4.2 with the ones introduced by Binder (1992) and Lin (2000).

The difficulties of defining "proper" longitudinal sampling weights were discussed in section 2.6. In the same spirit, questions regarding the choice between the different yearly longitudinal sampling weights available were raised following the discussion on SLIDRET in section 4.2. If one intends to use the proposed weighted variance estimators of section 2.6 or other weighted procedures to carry out inference from longitudinal studies, these issues regarding the choice of good longitudinal sampling weights must be researched further. This choice of weights will depend on which of the different assumptions made in chapter 3 (e.g., uninformative sampling design, independent and non-informative censoring) are not satisfied.

The simulation study of section 5.4.3 raised numerous questions. It would be interesting to try other numerical maximization techniques for the equilibrium scenario to see if they perform better than the S-Plus nlminb function. Even though variable metric methods (on which the S-Plus nlminb function is based) do not explicitly use first and second order derivatives, numerical approximation of these derivatives are made within the maximization algorithm. Other methods like the simplex and Powell's methods might yield better results. It would also be interesting to look at the estimation efficiency of other quantities than $\hat{\lambda}$ and $\hat{\beta}$. For example, $\hat{S}(t) = \exp\{(\hat{\lambda}t)^{\hat{\beta}}\}$ is likely to be of greater interest than $\hat{\lambda}$ or $\hat{\beta}$ on their own.

Appendix A

Appendix of chapter 4

A.1 Additional tables for section 4.2

Since table A.1 is too large, please see next page.

	$\hat{\boldsymbol{\beta}}$	$\sqrt{\widehat{V}_R(\hat{\beta})}$
	$n = 14,978$	$n = 14,978$
I(Sex)	-4.70×10^{-1}	(1.02×10^{-1})
I(Looked)	6.70×10^{-2}	(2.29×10^{-2})
I(Children)	-2.53×10^{-2}	(2.71×10^{-2})
Age	-2.23×10^{-2}	(2.26×10^{-3})
Education	-1.45×10^{-3}	(4.90×10^{-3})
Hourly-wage	1.70×10^{-2}	(2.48×10^{-3})
I(EI)	2.35×10^{-1}	(2.34×10^{-2})
I(Minority)	-1.35×10^{-1}	(1.00×10^{-1})
I(Winter)	3.07×10^{-1}	(3.62×10^{-2})
I(Sex)×Education	3.55×10^{-2}	(7.50×10^{-3})
I(Sex) × I(Children)	-1.11×10^{-1}	(4.04×10^{-2})
Age×I(Children)	2.44×10^{-2}	(1.73×10^{-3})
I(Sex)×Hourly-wage	-2.11×10^{-2}	(3.70×10^{-3})
I(Winter) × I(Looked)	-1.57×10^{-1}	(4.30×10^{-2})
Age×Hourly-wage	-1.20×10^{-3}	(1.44×10^{-4})
Age×I(EI)	1.53×10^{-2}	(1.81×10^{-3})

Note: s.d.'s computed using $\widehat{V}_R(\hat{\beta})$, given by (3.19).

Table A.1: Preliminary unweighted marginal Cox PH SLID data analysis for jobless spells.

$\hat{\boldsymbol{\beta}}$	$\sqrt{\widehat{V}_R(\hat{\beta})}$ $n = 14,978$	$\sqrt{\mathcal{I}(\hat{\beta})^{-1}}$ $n = 14,978$	$\sqrt{\widehat{V}_R(\hat{\beta})}$ $n = 16,682$	$\sqrt{\mathcal{I}(\hat{\beta})^{-1}}$ $n = 16,682$
I(Sex)	1.01×10^{-1}	8.83×10^{-2}	9.56×10^{-2}	8.43×10^{-2}
I(Looked)	2.50×10^{-2}	2.35×10^{-2}	2.37×10^{-2}	2.23×10^{-2}
I(Children)	2.71×10^{-2}	2.53×10^{-2}	2.56×10^{-2}	2.39×10^{-2}
Age	2.47×10^{-3}	2.17×10^{-3}	2.34×10^{-3}	2.06×10^{-3}
Education	4.91×10^{-3}	4.40×10^{-3}	4.64×10^{-3}	4.19×10^{-3}
Hourly-wage	2.44×10^{-3}	2.22×10^{-3}	2.32×10^{-3}	2.08×10^{-3}
I(EI)	2.51×10^{-2}	2.31×10^{-2}	2.34×10^{-2}	2.16×10^{-2}
I(Minority)	9.64×10^{-2}	8.90×10^{-2}	8.79×10^{-2}	8.20×10^{-2}
I(Winter)	3.50×10^{-2}	3.36×10^{-2}	3.29×10^{-2}	3.18×10^{-2}
Age^2	9.13×10^{-5}	7.92×10^{-5}	8.54×10^{-5}	7.48×10^{-5}
I(Winter) $\times$ I(Weeks > 52)	3.83×10^{-2}	5.34×10^{-2}	3.62×10^{-2}	5.09×10^{-2}
I(Sex)$\times$Education	7.38×10^{-3}	6.60×10^{-3}	6.96×10^{-3}	6.25×10^{-3}
I(Sex) $\times$ I(Children)	4.04×10^{-2}	3.73×10^{-2}	3.77×10^{-2}	3.50×10^{-2}
Age$\times I$(Children)	1.99×10^{-3}	1.80×10^{-3}	1.87×10^{-3}	1.71×10^{-3}
I(Sex)$\times$Hourly-wage	3.62×10^{-3}	3.40×10^{-3}	3.48×10^{-3}	3.25×10^{-3}
I(Winter) $\times$ I(Looked)	3.98×10^{-2}	4.02×10^{-2}	3.76×10^{-2}	3.81×10^{-2}
Age$\times$Hourly-wage	1.52×10^{-4}	1.34×10^{-4}	1.44×10^{-4}	1.29×10^{-4}
Age$\times I$(EI)	1.97×10^{-3}	1.76×10^{-3}	1.87×10^{-3}	1.68×10^{-3}

Table A.2: "Naive" and "robust" (with household ID as clusters) standard deviation estimates for the unweighted marginal Cox PH SLID data analyses for jobless spells.

Appendix B

Appendix of chapter 5

B.1 Algorithm for the simulations of section 5.3.2

Step 1: Generate $b_1, \ldots, b_n$ and $\tilde{t}_1, \ldots, \tilde{t}_n$ according to independent $\text{Exp}(\lambda)$ d.f.'s;

Step 2: Generate $c_1, \ldots, c_n$ according to independent $\text{Unif}(0, a)$ d.f.'s, where $a = 6.2\,(1 - \alpha)/\lambda$ and α is the expected percentage of right-censored observations (see justification below);

Step 3: If $\tilde{t}_i > c_i$, set $\tilde{t}_i = c_i$ and $\delta_i = 0$; otherwise, set $\delta_i = 1$;

Step 4: Compute $t_i = b_i + \tilde{t}_i$ for $i = 1, \ldots, n$;

Step 5: Compute $\hat{\lambda}_1$, $\hat{\lambda}_2$ and $\hat{\lambda}_3$ from the score functions give in section 5.3.1;

Step 6: Repeat steps 1 to 5 a 1000 times;

Step 7: Compute $\widehat{\widehat{\text{Var}}}(\hat{\lambda}_j)$, given by (5.26), for $j = 1, 2, 3$;

Step 8: Compute $\hat{\hat{\mathcal{I}}}(\hat{\lambda}_j)$'s (for $j = 1, 2, 3$) by inverting the quantities in step 7;

Step 9: Repeat steps 1 to 8 for $\lambda = 0.1, 0.2, 0.3, \ldots, 2$.

This algorithm was repeated for $n = 10$ and 100 using S-Plus.

Justification of the choice of a

Let $T \sim \text{Exp}(\lambda)$ and $C \sim \text{Unif}(0, a)$. The goal is to find a such that $\Pr\{C - T < a\} = \alpha$. To this end, first note that the joint d.f. of T and C is given by,

$$\begin{aligned} f_{T,C}(t,c) &= f_T(t)\, f_C(c) \qquad \text{since independent} \\ &= \lambda \exp\{-\lambda t\}\, \frac{1}{a} \qquad \text{for } 0 \le t \text{ and } 0 \le c \le a\ . \end{aligned}$$

Let $U = C - T$ and $V = T$; then,

$$J = \begin{vmatrix} \frac{\partial T}{\partial U} & \frac{\partial T}{\partial V} \\ \frac{\partial C}{\partial U} & \frac{\partial C}{\partial V} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 1 & 1 \end{vmatrix} = -1$$

Thus,

$$g_{U,V}(u,v) = \lambda \frac{1}{a} \exp\{-\lambda v\} \qquad \text{for } 0 \le v \text{ and } 0 \le u + v \le a$$

and

$$g_U(u) = \begin{cases} \frac{1}{a}(\exp\{\lambda u\} - \exp\{\lambda(u-a)\}) & \text{for } u \le 0 \\ \frac{1}{a}(1 - \exp\{\lambda(u-a)\}) & \text{for } 0 < u \le a\ . \end{cases}$$

Finally,

$$\Pr\{C - T < 0\} = \Pr\{U < 0\} = \int_{-\infty}^{0} \frac{1}{a} (\exp\{\lambda u\} - \exp\{\lambda(u-a)\})\, du = \frac{1}{a\lambda}(1 - \exp\{-\lambda a\})\ .$$

Taylor's expansion of $\exp\{x\}$, yields the following simplification,

$$\frac{1}{a\lambda}\,(1 - \exp\{-\lambda a\}) = \frac{1}{a\lambda} - \frac{1}{a\lambda}\left(1 - \lambda a + \frac{(\lambda a)^2}{2!} + \dots\right) \approx 1 - \frac{\lambda a}{2}\ .$$

Therefore, $\Pr\{C - T < 0\} \approx 1 - \lambda a/2 = \alpha$ and $a \approx 2(1-\alpha)/\lambda$. However, the factor 2 was too small to yield the proper percentage of right-censored observations and was increased to 6.2.

B.2 Computation of $[I(\lambda, \beta)]_{2,2}$ in (5.38)

From (5.37),

$$E\big([\mathcal{I}(\lambda,\beta)]_{2,2}\big) = n/\beta^2 + nE\Big((\lambda T_i)^\beta \ln^2(\lambda T_i)\Big) - nE\Big((\lambda B_i)^\beta \ln^2(\lambda B_i)\Big)\ , \qquad \text{(B.1)}$$

where

$$E\Big((\lambda T_i)^\beta \ln^2(\lambda T_i)\Big) = \int_0^\infty (\lambda u)^\beta \ln^2(\lambda u)\lambda\beta(\lambda u)^{\beta-1}\exp\{-(\lambda u)^\beta\}\,du\ . \tag{B.2}$$

Let $v = (\lambda u)^\beta$, then,

$$\begin{aligned} E\Big((\lambda T_i)^\beta \ln^2(\lambda T_i)\Big) &= \frac{1}{\beta^2}\int_0^\infty v\,(\ln v)^2\exp\{-v\}\,dv \\ &= \frac{1}{\beta^2}\Gamma''(2)\ , \end{aligned} \tag{B.3}$$

where, by definition,

$$\Gamma^{(k)}(x) = \frac{\partial^k}{\partial x^k}\Big\{\Gamma(x)\Big\} = \int_0^\infty u^{x-1}\,(\ln u)^k\,\exp\{-u\}\,du\ . \tag{B.4}$$

Now, by the properties of the gamma and digamma functions,

$$\Gamma''(2) = \frac{1}{\Gamma(2)}\big(\psi'(2)(\Gamma(2))^2 + (\psi(2)\Gamma(2))^2\big) = \gamma^2 - 2\gamma + \pi^2/6\ . \tag{B.5}$$

Combining (B.3) and (B.5) yields,

$$E\Big((\lambda T_i)^\beta \ln^2(\lambda T_i)\Big) = \frac{1}{\beta^2}(\gamma^2 - 2\gamma + \pi^2/6) \approx \frac{1}{\beta^2}0.8237\ , \tag{B.6}$$

for $i = 1, \ldots, n$.

From the same notions as the ones used in the calculation of (B.2),

$$E\Big((\lambda B_i)^\beta \ln^2(\lambda B_i)\Big) = \frac{1}{\beta^3}\big(\psi'(1/\beta) + 2\beta\psi(1/\beta) + \psi^2(1/\beta)\big)\ , \tag{B.7}$$

for $i = 1, \ldots, n$.

Finally, combining (B.1), (B.6) and (B.7) yields the second diagonal element of (5.38); that is,

$$[I(\lambda,\beta)]_{2,2} = \frac{n}{\beta^2}(1 + \gamma^2 - 2\gamma + \pi^2/6) - \frac{n}{\beta^3}\big(\psi'(1/\beta) + 2\beta\psi(1/\beta) + \psi^2(1/\beta)\big)\ .$$

B.3 Properties of the family of gamma functions

A special value of the digamma function is,

$$\psi(1) = -\gamma \ , \tag{B.8}$$

where $\gamma \approx 0.5772157$ is the Euler-Mascheroni constant.

Recurrence formula for the digamma and trigamma functions are, respectively,

$$\psi(x+1) = \psi(x) + 1/x \tag{B.9}$$

and

$$\psi'(x+1) = \psi'(x) - 1/x^2 \ . \tag{B.10}$$

The digamma and trigamma functions can be express in terms of the gamma function and its derivatives; that is,

$$\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)} \tag{B.11}$$

$$\psi'(x) = \frac{\Gamma''(x)\Gamma(x) - (\Gamma'(x))^2}{(\Gamma(x))^2} \ . \tag{B.12}$$

Similarly, the derivative of the incomplete gamma function can be express as

$$\begin{aligned} \frac{\partial}{\partial y}\Big\{\Gamma(x, g(y))\Big\} &= \int_{g(y)}^{\infty} u^{x-1} \exp\{-u\}\, du \\ &= -g(y)^{x-1} \exp\{-g(y)\}\, g'(y) \ . \end{aligned} \tag{B.13}$$

See Weisstein (1999) and Abramowitz & Stegun (1965) for further references on the gamma, digamma, trigamma and incomplete gamma functions.

Bibliography

Aalen, O. O. & Husebye, E. (1991), 'Statistical analysis of repeated events forming renewal processes', *Statistics in Medicine* **10**, 1227–1240.

Abramowitz, M. & Stegun, I. A., eds (1965), *Handbook of Mathematical Functions*, Dover Publications, New York.

Allison, P. D. (1984), *Event History Analysis, Regression for Longitudinal Event Data*, Sage, Beverly Hills.

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1982), 'Linear nonparametric tests for comparison of counting processes, with applications to censored survival data', *International Statistical Review* **50**, 219–258.

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993), *Statistical Models Based on Counting Processes*, Springer – Verlag, New York.

Asgharian, M., M'Lan, C. E. & Wolfson, D. B. (2002), 'Length-biased sampling with right censoring: An unconditional approach', *Journal of the American Statistical Association* **97**(457), 201–209.

Binder, D. A. (1983), 'On the variances of asymptotically normal estimators from complex surveys', *International Statistical Review* **51**(3), 279–292.

Binder, D. A. (1992), 'Fitting Cox's proportional hazards models from survey data', *Biometrika* **79**(1), 139–147.

Binder, D. A. (1998), 'Longitudinal surveys: Why are these surveys different from all other surveys?', *Survey Methodology* **24**(2), 101–108.

Binder, D. A. & Roberts, G. R. (2001), 'Can informative designs be ignorable?', Survey Research Methods Section Newsletter Issue 12, American Statistical Association.

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975), *Discrete Multivariate Analysis*, The Massachusetts Institute of Technology Press, Cambridge.

Boudreau, C. & Lawless, J. F. (2001), 'Survival analysis based on the proportional hazards model and survey data', Working Paper #2001–10, Dept. of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario.

Citro, C., Hernandez, D. & Herriot, R. (1986), Longitudinal household concepts in SIPP: Preliminary results, *in* 'Proceedings of the Bureau of the Census' Second Annual Research Conference', pp. 598–611.

Cnaan, A. & Ryan, L. (1989), 'Survival analysis in natural history studies of disease', *Statistics in Medicine* **8**(10), 1255–1268.

Cochran, W. G. (1977), *Sampling Techniques*, 3 edn, John Wiley & Sons, New York.

Coder, J., Burkhead, D., Feldman-Harkins, A. & McNeil, J. (1987), 'Preliminary data from the SIPP 1983–84 longitudinal research file', SIPP Working Paper No. 8702. U.S. Bureau of the Census, Washington, DC.

Cotton, C. & Giles, P. (1998), 'The seam effect in the Survey of Labour and Income Dynamics', SLID Working Paper No. 75F0002M. Statistics Canada, Ottawa, Ontario.

Cox, D. R. (1969), Some sampling problems in technology, *in* N. L. Johnson & H. J. Smith, eds, 'New Developments in Survey Sampling', John Wiley & Sons, New York, pp. 506–527.

Cox, D. R. (1972), 'Regression models and life-tables (with discussion)', *Journal of the Royal Statistical Society, Series B — Methodological* **34**, 187–220.

Cox, D. R. & Oakes, D. (1984), *Analysis of Survival Data*, Chapman & Hall, London.

Deming, W. E. (1950), *Some Theory of Sampling*, Dover, New York.

Ernst, L. R. (1989), Weighting issues for longitudinal household family estimates, *in* Kasprzyk et al. (1989), pp. 139–159.

Fienberg, S. E. (1989), Modeling considerations: Discussion from a modeling perspective, *in* Kasprzyk et al. (1989), pp. 566–574.

Fitzmaurice, G. M., Heath, A. F. & Cox, D. R. (1997), 'Detecting overdispersion in large scale surveys: Applications to a study of education and social class in Britain', *Applied Statistics* **46**, 415–432.

Fleming, T. R. & Harrington, D. P. (1991), *Counting Processes and Survival Analysis*, John Wiley & Sons, New York.

Folsom, R., LaVange, L. & Williams, R. L. (1989), A probability sampling perspective on panel data analysis, *in* Kasprzyk et al. (1989), pp. 108—138.

Freedman, D., Thornton, A., Camburn, D., Alwin, D. & Young-DeMarco, L. (1988), The life history calendar: a technique for collecting retrospective data, *in* C. Clogg, ed., 'Sociological Methodology 1988', American Sociological Association, Washington, DC, pp. 37–68.

Goldstein, H. (1979), *The Design and Analysis of Longitudinal Studies*, Academic Press, London.

Grambsch, P. M. & Therneau, T. M. (1994), 'Proportional hazards tests and diagnostics based on weighted residuals', *Biometrika* **81**(3), 515–526.

Grimmett, G. R. & Stirzaker, D. R. (1992), *Probability and Random Processes*, 2 edn, Clarendon Press, Oxford, England.

Groves, R. M. (1989), *Survey Errors and Survey Costs*, John Wiley & Sons, New York.

Guo, G. (1993), 'Event-history analysis for left-truncated data', *Sociological Methodology* **23**, 217–243.

Hald, A. (1952), *Statistical Theory with Engineering Application*, John Wiley & Sons, New York.

Heckman, J. J. & Singer, B. (1986), Econometric analysis of longitudinal data, *in* Z. Griliches & M. D. Intriligator, eds, 'Handbook of Econometrics', Vol. III, Elsevier Science Publisher, chapter 29, pp. 1689–1763.

Hoem, J. M. (1989), The issue of weights in panel surveys of individual behavior, *in* Kasprzyk et al. (1989), pp. 539–565.

Hyde, J. (1977), 'Testing survival under right censoring and left truncation', *Biometrika* **64**, 225–230.

Hyde, J. (1980), Survival analysis with incomplete observations, *in* R. G. J. Miller, B. Efron, B. W. J. Brown & L. E. Moses, eds, 'Biostatistics Casebook', John Wiley & Sons, New York, pp. 31–46.

Institute for Social Research (1984), *User Guide to the Panel Study of Income Dynamics*, Survey Research Center, University of Michigan, Ann Arbor, MI.

Jean, A. C. & McArthur, E. K. (1987), 'Tracking persons over time', SIPP Working Paper No. 8701. U.S. Bureau of the Census, Washington, DC.

Kalbfleisch, J. D. & Lawless, J. F. (1991), 'Regression models for right truncated data with applications to AIDS incubation times and reporting lags', *Statistica Sinica* **1**, 19–32.

Kalbfleisch, J. D. & Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data*, 2 edn, John Wiley & Sons, Hoboken, NJ.

Kalton, G. (1989), Modeling considerations: Discussion from a survey sampling perspective, *in* Kasprzyk et al. (1989), pp. 575—585.

Kalton, G. & Brick, J. M. (1995), 'Weighting schemes for household panel surveys', *Survey Methodology* **21**(2), 33–34.

Kalton, G., Miller, D. P. & Lepkowski, J. (1992), Analyzing spells of program participation in the SIPP, Technical report, Survey Research Center, University of Michigan.

Kalton, G. & Miller, M. E. (1991), 'The seam effect with Social Security income in the Survey of Income and Program Participation', *Journal of Official Statistics* **7**, 235–245.

Kasprzyk, D., Duncan, G., Kalton, G. & Singh, M. P., eds (1989), *Panel Surveys*, John Wiley & Sons, New York.

Keiding, N. (1992), Independent delayed entry, *in* Klein & Goel (1992), pp. 309–326.

Klein, J. P. & Goel, P. K., eds (1992), *Survival Analysis: State of the Art*, Vol. 211 of *ASI Serie E: Applied Sciences*, NATO ASI Series, Kluwer Academic Publishers, Dordrecht.

Klein, J. P. & Moeschberger, M. L. (1997), *Survival Analysis: Techniques for Censored and Truncated Data*, Springer-Verlag, New York.

Korn, E. L. & Graubard, B. I. (1999), *Analysis of Health Surveys*, John Wiley & Sons, New York.

Lawless, J. F. (2003), *Statistical Models and Methods for Lifetime Data*, 2 edn, John Wiley & Sons, Hoboken, NJ.

Lawless, J. F. (2003b), Event history analysis and longitudinal surveys, *in* R. Chambers & C. J. Skinner, eds, 'Analysis of Survey Data', John Wiley & Sons, New York.

Lawless, J. F. & Boudreau, C. (2002), Modelling and analysis of duration data from longitudinal surveys, *in* 'Proceedings of the XIX International Methodology Symposium', Statistics Canada, Ottawa, Ontario.

Lee, E. W., Wei, L. J. & Amato, D. A. (1992), Cox-type regression analysis for large numbers of small groups of correlated failure time observations, *in* Klein & Goel (1992), pp. 237–247.

Leiderman, P. H., Babu, B., Kagia, J., Kraemer, H. C. & Leiderman, G. F. (1973), 'African infant precocity and some social influences during the first year', *Nature* **242**, 247–249.

Lemaître, G. (1992), 'Dealing with the seam problem for the Survey of Labour and Income Dynamics', SLID Research Paper 92-05. Statistics Canada, Ottawa, Ontario.

Liang, K.-Y., Self, S. G. & Chang, Y.-C. (1993), 'Modelling marginal hazards in multivariate failure time data', *Journal of the Royal Statistical Society, Series B — Methodological* **55**(2), 441–453.

Lillard, L. A. (1989), Sample dynamics: Some behavioral issues, *in* Kasprzyk et al. (1989), pp. 497–511.

Lin, D. Y. (2000), 'On fitting Cox's proportional hazards models to survey data', *Biometrika* **87**(1), 37–47.

Lin, D. Y. & Wei, L. J. (1989), 'The robust inference for the Cox proportional hazards model', *Journal of the American Statistical Association* **84**(408), 1074–1078.

Lin, D. Y., Wei, L.-J., Yang, I. & Ying, Z. (2000), 'Semiparametric regression for the mean and rate functions of recurrent events', *Journal of the Royal Statistical Society, Series B — Methodological* **62**, 711–730.

Little, R. J. A. (1989), Survey inference with weights for differential sample selection or nonresponse, *in* 'Proceedings of the Section on Survey Research Methods', American Statistical Association, pp. 62–69.

Michaud, S. & Webber, M. (1994), 'Measuring non-response in a longitudinal survey: the experience of the Survey of Labour and Income Dynamics', SLID Research Paper 94-16. Statistics Canada, Ottawa, Ontario.

Patterson, B. H., Dayton, C. M. & Graubard, B. I. (2002), 'Latent class analysis of complex sample survey data: Application to dietary data', *Journal of the American Statistical Association* **97**(459), 721–741.

Pfeffermann, D. (1993), 'The role of sampling weights when modeling survey data', *International Statistical Review* **61**(2), 317–337.

Pfeffermann, D. & Smith, T. M. F. (1985), 'Regression models for grouped populations in cross-section surveys', *International Statistical Review* **53**(1), 37–59.

Pollard, D. (1990), *Empirical Processes: Theory and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics, Hayward, CA.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (2002), *Numerical Recipes in C++: The Art of Scientific Computing*, 2 edn, Cambridge University Press, Cambridge.

Rao, C. R. (1973), *Linear Statistical Inference and its Applications*, 2 edn, John Wiley & Sons, New York.

Rebolledo, R. (1980), 'Central limit theorems for local martingales', *Z. Wahrsch. Verw. Gebiete* **51**(3), 269–286.

Research Triangle Institute (2001), *SUDAAN User's Manual, Release 8.0*, Research Triangle Park, NC.

Rockafellar, R. T. (1970), *Convex Analysis*, Princeton mathematical series #28, Princeton University Press, Princeton, NJ.

Ross, S. M. (1996), *Stochastic Processes*, 2 edn, John Wiley & Sons, New York.

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* **63**(3), 581–592.

Särndal, C.-E., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer – Verlag, New York.

Schoen, R. (1975), 'Constructing increment-decrement life tables', *Demography* **12**, 313–324.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New York.

Skinner, C. J., Holt, D. & Smith, T. M. F., eds (1989), *Analysis of Complex Surveys*, John Wiley & Sons, Chichester.

Snow, J. (1855), *On the Model of Communication of Cholera*, 2 edn, Churchill.

Spiekerman, C. F. & Lin, D. Y. (1998), 'Marginal regression models for multivariate failure time data', *Journal of the American Statistical Association* **93**(443), 1164–1175.

Statistics Canada (1997), *Survey of Labour and Income Dynamics Microdata User's Guide*, Ottawa, Ontario. Catalogue 75M0001GPE.

Sugden, R. A. & Smith, T. M. F. (1984), 'Ignorable and informative designs in survey sampling inference', *Biometrika* **71**(3), 495–506.

Therneau, T. M. & Grambsch, P. M. (2000), *Modeling Survival Data: Extending the Cox Model*, Springer – Verlag, New York.

Thompson, Jr, W. A. (1977), 'On the treatment of grouped observations in life studies', *Biometrics* **33**, 463–470.

Thompson, M. E. (1997), *Theory of Sample Surveys*, Chapman & Hall, London.

Tsai, W.-Y. (1990), 'Testing the assumption of independence of truncation time and failure time', *Biometrika* **77**(1), 169–177.

Tuma, N. B. & Hannan, M. T. (1984), *Social Dynamics: Models and Methods*, Academic Press, New York.

Turnbull, B. W. (1974), 'Nonparametric estimation of a survivorship function with doubly censored data', *Journal of the American Statistical Association* **69**(345), 169–173.

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

U.S. Census Bureau (2001), *Survey of Income and Program Participation Users' Guide*, 3 edn, Washington, D.C.

van der Vaart, A. W. & Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Springer – Verlag, New York.

Vardi, Y. (1989), 'Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation', *Biometrika* **76**(4), 751–761.

Wei, L. J., Lin, D. Y. & Weissfeld, L. (1989), 'Regression analysis of multivariate incomplete failure time data by modeling marginal distributions', *Journal of the American Statistical Association* **84**(408), 1065–1073.

Weisstein, E. W. (1999), *CRC Concise Encyclopedia of Mathematics*, CRC Press, Boca Raton, FL.

Williams, R. L. (1995), 'Product-limit survival functions with correlated survival times', *Lifetime Data Analysis* **1**(2), 171–186.

Yamaguchi, K. (1991), *Event History Analysis*, Sage Publications, Newbury Park, CA.

Young, N. (1989), 'Wave seam effects in the SIPP', SIPP Working Paper No. 8921. U.S. Bureau of the Census, Washington, DC.