Analysis of duration data from longitudinal surveys subject to loss to follow-up

by

C. Dagmar Mariaca Hajducek

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, 2010

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

Data from longitudinal surveys give rise to many statistical challenges. They often come from a vast, heterogeneous population and from a complex sampling design. Further, they are usually collected retrospectively at intermittent interviews spaced over a long period of time, which gives rise to missing information and loss to follow-up. As a result, duration data from this kind of surveys are subject to dependent censoring, which needs to be taken into account to prevent biased analysis. Methods for point and variance estimation are developed using Inverse Probability of Censoring (IPC) weights. These methods account for the random nature of the IPC weights and can be applied in the analysis of duration data in survey and non-survey settings. The IPC estimation techniques are based on parametric estimating function theory and involve the estimation of dropout models. Survival distributions without covariates are estimated via a weighted Kaplan-Meier method and regression modeling through the Cox Proportional Hazards model and other models is based on weighted estimating functions. The observational frameworks from Statistics Canada's Survey of Labour and Income Dynamics (SLID) and the UK Millenium Cohort Study are used as motivation, and durations of jobless spells from SLID are analyzed as an illustration of the methodology. Issues regarding missing information from longitudinal surveys are also discussed.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Jerald F. Lawless, for his patience, encouragement, constructive advise, and for sharing with me his knowledge and experience.

I am grateful to my internal examiners, Dr. Mary E. Thompson and Dr. Christian Boudreau, for their valuable suggestions and comments. Many thanks to my external examiners, Dr. John Goyder and Dr. Milorad Kovacevic for taking the time to read and assess this thesis, and for their useful comments. I would also like to thank Dr. Georgia Roberts for her guidance with regard to data from SLID, the survey that was used for illustration and which motivated the discussion on features of longitudinal surveys. Her kindness and encouragement are greatly appreciated. Sincere thanks are extended to Dr. Pat Newcombe Welch from the South-Western Ontario Research Data Centre (SWORDC) for her timely assistance.

My gratitude also goes to the National Council for Science and Technology of Mexico (CONACyT) for awarding me the scholarship that sustained me while completing an important part of this thesis.

Many thanks to MITACS and Statistics Canada for providing me with the opportunity of spending a four month internship at Statistics Canada during the development of this work. I am thankful to the South-Western Ontario Research Data Centre (SWORDC) for providing me access to the data from SLID in Waterloo.

I would also like to thank my husband for his support, love, and patience, which have been crucial for this accomplishment. I am deeply grateful to my mother for her loving advise and encouragement, and for her always positive example. I would like to thank my friends inside and outside of UW, since they have been an important part of my academic and personal life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Settings of interest

A life history process is usually characterized by events that are experienced by individuals through their lifetime pertaining to health, education, labor experience, social dynamics, economic history, etc. These kind of processes may often be represented by a set of states and the transitions an individual may experience among them.

In studying a certain disease characterized by the states "infected" and "not infected", there may be interest in analyzing the time to infection or the times between infections in patients and the variables that affect them such as a treatment or physiological characteristics. When examining the events that occur in an individual's employment history characterized by states such as being "out of the labor force", "employed" and "unemployed", one may be interested in examining the time to experiencing one of these states or the length of a sojourn in a particular state, and its relationship with variables like age, gender, education level, etc. It may also be of interest to estimate the distribution of the durations of the experienced jobless spells without considering covariates.

Data on life history processes are collected over time and may include specific

information on the timing and duration of events. Information can be collected from cohorts that are randomly selected, from observational studies on a population or cohorts from a population, or through longitudinal surveys. Prospective data are usually collected in intermittent interviews over a long period of time. Further, it often comes from a heterogeneous population, and the sampling scheme may involve clustering, stratification and unequal probabilities of selection. For instance, the longitudinal Survey of Labour and Income Dynamics (SLID) interviews Canadian individuals once a year, over periods of six years. Each year, SLID collects information about individual labour history, family composition and economic experience pertaining to the previous year. Since this survey includes individuals from across Canada, it relies on a complex sampling scheme that takes into account the heterogeneity of the population. Information about SLID can be found online at www.statcan.gc.ca/pub/75f0011x/4060256-eng.htm.

Another example of longitudinal surveys is the Millennium Cohort Study (MCS), providing data from children growing up in the four countries of the United Kingdom. This is a complex survey that aims at understanding the social and economic conditions surrounding birth and early childhood and collects information regarding the develpment of children that were born in 2000 and 2001. Information is collected from children at ages 9 months, 3, 5 and 7 years old. Online information about the MCS can be found at www.cls.ioe.ac.uk/text.asp?section=000100020001.

When data are collected over spaced interviews and over a long period of time, it is usually the case that information is lost partially or even completely. In some cases the individual may have been contacted, but the information was not collected in its totality. In some other cases, individuals may be lost to follow-up at some time before the end of the study and no further information is collected at all. Our attention will focus on the latter scenario, where it is often reasonable to assume that the loss to follow-up mechanism is related to the life history of individuals, that is, to the events they experience and to covariates.

In the above examples, loss to follow-up becomes substantial over time. SLID

samples are typically in the 25-30 percent range of loss to follow-up by the end of the six years. In the MCS study, there was a loss to follow-up rate of 28 percent in the first wave and 42 percent by the second wave (Plewis,[47]). Dependent loss to follow-up has been considered by many authors in the context of continuous and binary outcomes (e.g. Robins et al., [51]; Miller et al., [44]; Preisser et al., [48]); however, event history or duration analysis where data are collected retrospectively at each interview time has not been considered.

Methodology for the analysis of durations can be applied in this kind of setting. In general, it is used to examine the times to events, the times between events or the lengths of sojourns in states. In the simplest case, the time to occurrence of only one event per individual can be analyzed through standard survival analysis. Duration analysis is also used for the analysis of the times between successive events experienced by the same subject, that is, of multiple durations per individual. Even though it has been thoroughly studied in many settings, the analysis of durations has not yet received much attention under the assumption of dependent loss to follow-up and in particular in the context of complex survey data.

This chapter provides the theoretical basis for the methods to be developed in this dissertation. Section 1.2 gives a discussion of survival analysis theory. Multistate models are discussed in section 1.3. Section 1.4 introduces the statistical challenges when analyzing survey data. Section 1.5 presents a summary of survival methods extended to survey data that are found in the literature. These methods involve single durations per individuals and do not consider the issue of dependent loss to follow-up. Section 1.6 presents additional features of longitudinal surveys and describes dependent loss to follow-up. Finally, section 1.7 provides an outline of the remainder of the thesis.

1.2 Survival analysis

A failure time is defined as a duration, survival time or time between two events of particular interest. Survival analysis consists of the study of failure times taking into consideration their relation with other variables and with censoring processes. We will be focusing on the right type of censoring in which the individual was not observed to fail during a follow-up period.

Likelihood function for right censored data

An individual's lifetime or failure time is denoted by T_i and the censoring time is represented by C_i. In dealing particularly with a continuous time specification, we let $f(t_i)$ and $g(t_i)$ be the density functions of the failure and censoring times for individual i, $i = 1, 2, ..., n$ respectively. Further, let $S(t_i) = P(T_i > t_i)$ and $G(t_i) = P(C_i > t_i)$ denote the survivor functions of T_i and C_i. Also, let $h(t) = f(t)/S(t)$ be the hazard function, that describes the instantaneous rate of failure at time t. The observed time is represented by $t_i = \min(T_i, C_i)$ and the status indicator $\delta_i = I(T_i \leq C_i)$.

If T_i and C_i are independent, the likelihood function for right censored data $\{(t_i, \delta_i) \quad i = 1, 2, ..., n\}$ is given by:

$$\begin{aligned} L &= \prod_{i=1}^{n} \left[f(t_i)G(t_i)\right]^{\delta_i} \left[S(t_i)g(t_i)\right]^{1-\delta_i} \\ &\propto \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} = \prod_{i=1}^{n} h(t_i)^{\delta_i} \exp\left\{-\int_0^{t_i} h(u)du\right\}. \end{aligned} \tag{1.1}$$

This likelihood (1.1) can be extended to include covariates, and parametric regression models can be examined. Among the models that can be used to specify the density and hazard functions in (1.1), this work will focus on the the log-location-scale and the proportional hazards regression models.

The solution $\hat{\theta}$ to the score equation $U(\theta) = \partial log(L(\theta))/\partial\theta = 0$ where θ is a parameter of dimension p, in most cases maximizes $L(\theta)$ and is called the

maximum likelihood estimate (MLE). Tests and interval estimates for θ can be performed using the large-sample approximation of a p-variate normal distribution of the estimators $\hat{\theta}$. That is, $\sqrt{n}(\hat{\theta} - \theta)$ is asymptotically Normally distributed with zero mean and variance V, which is estimated by $\hat{V} = I(\hat{\theta})^{-1}$, where $I(\theta) = -\partial^2 \log L(\theta)/\partial\theta\partial\theta'$. For a more detailed discussion, refer to Kalbfleisch and Prentice [25] and Lawless [32].

It is important to note that (1.1) is based on the joint distribution for the censoring and failure times. The main assumptions for (1.1) to be valid are that (i) the failure times for individuals occur independently; (ii) failure and censoring times are independent given covariates in the failure time model; and for the right hand side of (1.1), that (iii) the distribution of the censoring times does not involve parameters that specify the failure times distribution (noninformative censoring).

The assumptions (ii) and (iii) can be relaxed somewhat and the expression in the right hand side of (1.1) is no longer a likelihood but is regarded as a partial likelihood. For a detailed discussion about likelihood and partial likelihood estimation with lifetime data under independent and noninformative censoring see Lawless ([32], p.59-60).

Non-parametric estimation in the absence of covariates.

The non-parametric estimator of the survivor function, known as the Product Limit or Kaplan-Meier (KM) estimator, is a function of the proportion of failure times d_t and the number of at risk individuals n_t at each time t, giving $\hat{h}(t) = d_t/n_t$. These two quantities are expressed as $d_t = \sum_{i=1}^{n} I(t_i = t, \delta_i = 1)$, $n_t = \sum_{i=1}^{n} I(t_i \geq t)$, where δ_i is the censoring indicator introduced earlier. The K-M estimate has the following form:

$$\hat{S}(t) = \prod_{s \leq t} [1 - \hat{h}(s)]. \tag{1.2}$$

It is understood that $\hat{h}(s) = 0$ whenever $d_s = 0$, $n_s > 0$ and is undefined when $n_s = 0$. The KM estimate can be derived as a non-parametric maximum likelihood

estimator of the discrete time formulation of the survivor function by constructing the likelihood of the lifetimes in terms of the hazard function $h(t)$ as the parameter of interest. For a detailed derivation, see Lawless [32], p.83.

As a maximum likelihood estimator, the asymptotic variance of the KM estimator can be obtained from standard maximum likelihood large sample theory. Let $t_1, t_2, ..., t_k$ be the distinct failure times in a sample. Then the asymptotic variance of the KM estimate is estimated by (Greenwood Formula):

$$\widehat{Var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}. \tag{1.3}$$

The nonparametric maximum likelihood estimator of the cumulative hazard function $H(t) = \int_0^t h(s)ds$ is the Nelson-Aalen (NA) estimator,

$$\hat{H}(t) = \sum_{j:t_j \leq t} d_j/n_j. \tag{1.4}$$

Since the K-M and N-A estimates are maximum likelihood estimates, in the discrete time case they both have large sample properties such as aymptotic normality and consistency, allowing for estimation of confidence intervals of survival probabilities and hypothesis tests using (empirical) likelihood ratio statistics. These results also extend to the case where $S(t)$ is continuous. The SPlus functions `survfit` and `kaplanMeier`, and the R function `survfit` can be used to obtain the K-M and N-A estimates. Analogously, the procedure `LIFETEST` from SAS is available. The SYSTAT/MYSTAT packages have the SURVIVAL module in which Kaplan-Meier estimates can be obtained. In the case of survey data, the SUDAAN package provides the function `KAPMEIER`, and the STATA package through the option `STS`.

Parametric regression models.

As mentioned earlier, explanatory variables for failure times may be included parametrically in (1.1). The discussion below involves a vector of fixed covariates x,

however, it can be generalized to external time varying covariates, which are not affected by the survival status.

The most widely used parametric regression models are those from the log-location-scale and the (parametric) proportional hazards (PH-also called relative risk) families. The log-location-scale models are usually specified in terms of the log-failure time $Y = \log(T)$, through the survivor function of $W = (Y - u(x))/b$, that is,

$$S(y|x) = S_W\left(\frac{y - u(x)}{b}\right), \tag{1.5}$$

where b is a scale parameter and the location parameter has usually the form $u(x) = x'\beta$, a function of the covariates x and β, px1 vectors. The most convenient way to express the relationship between Y and W is through the linear form $Y = u(x)+bW$, where the variable W is commonly distributed as standard normal, extreme value or logistic, and correspondingly, T is distributed as log-normal, Weibull or log-logistic.

Estimation and inference is performed based on the likelihood in (1.1) and the usual maximum likelihood asymptotic theory. When $u(x) = x'\beta$, we have $\theta = (\beta, b)$ and expression (1.1) becomes:

$$L(\beta, b) = \prod_{i=1}^{n}\left[\frac{1}{b}f_W\left(\frac{y_i - x_i'\beta}{b}\right)\right]^{\delta_i} S_W\left(\frac{y_i - x_i'\beta}{b}\right)^{1-\delta_i}, \tag{1.6}$$

where $y_i = \min\{Y_i, \log C_i\}$.The PH regression class of models is specified through the hazard function. The fully parametric PH models consist of a parametric baseline hazard $h_0(t;\eta)$ and some function of the covariates, usually $r(x) = \exp(x'\beta)$:

$$h(t|x) = h_0(t;\eta)r(x). \tag{1.7}$$

The likelihood for the model in (1.7) in terms of the hazard function has the form:

$$L(\eta, \beta) = \prod_{i=1}^{n}[h_0(t_i;\eta)\exp(x_i'\beta)]^{\delta_i}\exp\left\{-H_0(t_i;\eta)\mathrm{e}^{x_i'\beta}\right\}.$$

Estimation and inference for hazard based modelling can be readily implemented using optimization software based on the likelihood function above, as well as computation of the estimated asymptotic covariance matrix of the maximum likelihood estimators $(\hat{\eta}, \hat{\beta})$.

Statistical software for lifetime data is widely available. Among the parametric survival analysis software packages are SPlus/R through the function `censorReg`/ `survreg` and SAS through the `LIFEREG` procedure. References to other packages can be found in Lawless ([32], p.40).

Semi-parametric regression models.

The semi-parametric proportional hazards model introduced by Cox [17] has gained much popularity because it does not require a full parametric specification of the survival regression model, that is, the regression parameters in β can be estimated without specifying explicitly the baseline hazard $h_0(t)$. The only assumption required is the multiplicative relation of the covariates with the baseline hazard, though it still needs to be validated. With the usual covariate function $r(x) = \exp(x'\beta)$, the expression for the hazard function in (1.7) becomes:

$$h(t|x) = h_0(t)\exp(x'\beta). \tag{1.8}$$

Estimation for semi-parametric PH models is related to the concept of partial likelihood, first introduced by Cox [18]. It is based on the conditional probability that a given individual has a failure at time t, given that a failure actually has occurred at time t and the set of individuals at risk, $R(t)$ (individuals that have not failed and are uncensored at time t). The partial likelihood to estimate β from individuals $i = 1, 2, ..., n$ is

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{e^{\beta' x_i}}{\sum_{l=1}^{n} Y_l(t_i) e^{\beta' x_l}} \right)^{\delta_i}, \tag{1.9}$$

where $Y_i(t) = I(t_i \geq t)$ indicates whether individual i is at risk at time t and δ_i is the censoring indicator defined earlier. The expression in (1.9) is not an ordinary likelihood; however, in many applications it can still be treated as one. For details, refer to Lawless ([32], pp.349,551) and Kalbfleisch and Prentice ([25], p.99).

The score $U(\beta)$ and the information matrix $I(\beta)$ can be computed and used in optimization algorithms for estimation, likelihood ratio tests can be performed and

$\hat{\beta}$ is asymptotically normal with expectation β and covariance matrix estimated by $I(\hat{\beta})^{-1}$. Taking the logarithm of (1.9) and differentiating with respect to β, the px1 score vector may be obtained as:

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left[x_i - \frac{S^{(1)}(t_i, \beta)}{S^{(0)}(t_i, \beta)} \right], \quad \text{where} \tag{1.10}$$

$$S^{(0)}(t, \beta) = \sum_{i=1}^{n} Y_i(t) \exp(x_i' \beta) \qquad \text{and}$$

$$S^{(1)}(t, \beta) = \sum_{i=1}^{n} Y_i(t) x_i \exp(x_i' \beta).$$

When estimating the survivor function $S(t|x) = S_0(t)^{\exp(x'\beta)}$, regression coefficients are estimated from maximizing (1.9) and the baseline survivor function can be estimated using the Breslow or generalized Nelson-Aalen estimate $\hat{H}_0(t)$ and the relation $\hat{S}_0(t) = \exp(-\hat{H}_0(t))$. The estimate $\hat{H}_0(t)$ has the form (for details, see Lawless [32], ch.7):

$$\hat{H}_0(t) = \sum_{j:t_j \le t} \left\{ \frac{\delta_j}{\sum_{i=1}^{n} Y_i(t_j) e^{x_i' \hat{\beta}}} \right\} = \sum_{j:t_j \le t} \left\{ \frac{\delta_j}{S^{(0)}(t_j, \hat{\beta})} \right\}. \tag{1.11}$$

The asymptotic variance of $\hat{H}_0(t)$ is estimated by:

$$\widehat{Var}[\hat{H}_0(t)] = \sum_{i:t_i \le t} \frac{\delta_i}{S^{(0)}(t_i, \hat{\beta})^2} + \hat{W}(t)' I(\hat{\beta}) \hat{W}(t), \text{ where} \tag{1.12}$$

$$\hat{W}(t) = \sum_{i:t_i \le t} \frac{\delta_i \bar{x}(t_i, \hat{\beta})}{S^{(0)}(t_i, \hat{\beta})}, \quad \bar{x}(t, \beta) = \sum_{i=1}^{n} \frac{Y_i(t) x_i \exp(x_i' \beta)}{\sum_{l=1}^{n} Y_l(t) \exp(x_l' \beta)},$$

and $S^{(0)}(t, \hat{\beta})$ as in (1.10).

Sometimes it is desirable to define separate hazard functions representing the strata from a population, when it is assumed that individuals in the same stratum have proportional hazard functions, but not so for those in different strata. The stratified model with hazard function $h_j(t|x) = h_{0j}(t) e^{x'\beta}$ for stratum j, $j = 1, 2, ..., J$ is often used. Then the likelihood function is constructed as the product of the stratum-specific likelihoods $L_j(\beta)$ of the form (1.9) and the score and information functions described earlier are summed over the strata.

The methodology described above can be used in the case of time varying covariates. It is straightforward to replace the fixed covariate vector x by the time dependent covariate vector $x(t)$. The model in (1.8) can be extended to $h(t|x(t)) = h_0(t)\exp(x(t)'\beta)$ and the partial likelihood (1.9) as well as the score in (1.10) can be expressed similarly. Direct generalizations of (1.11) and (1.12) can also be made.

Software packages for the semi-parametric analysis just described include SPlus/R and SAS. The SPlus/R function `coxph` fits a Cox PH regression model with fixed or time dependent covariates and can handle stratification. Also, the function `cox.zph` in SPlus/R provides a test for the proportional hazards assumption. Analogously, SAS provides the procedure `PHREG`. A package that accommodates this type of estimation for survey data is SUDAAN with the function `SURVIVAL`.

1.3 Multistate models and consecutive durations

The survival theory described in the preceding section involves the time to a single event which can also be viewed as the time spent in one life state before making a transition to another. In a multistate process an individual is assumed to occupy one of a defined set of states $\{1, 2, ..., K\}$ at any given time. Multistate processes generate multiple lifetime variables per subject. The related lifetime variables may indicate the time at which each state was visited or the length of the sojourn in each state, for example.

One example of multistate processes is given by a study on patients treated for colon cancer (Moertel et al., [45]). The states under study are "treated and disease free","disease recurrence" and "death". Related times to events are: time from treatment to disease recurrence and the times to death from treatment or from the recurrence of the disease.

Labour studies provide one more example that involves the analysis of transitions between a set of states. These studies involve the states "employed", "un-

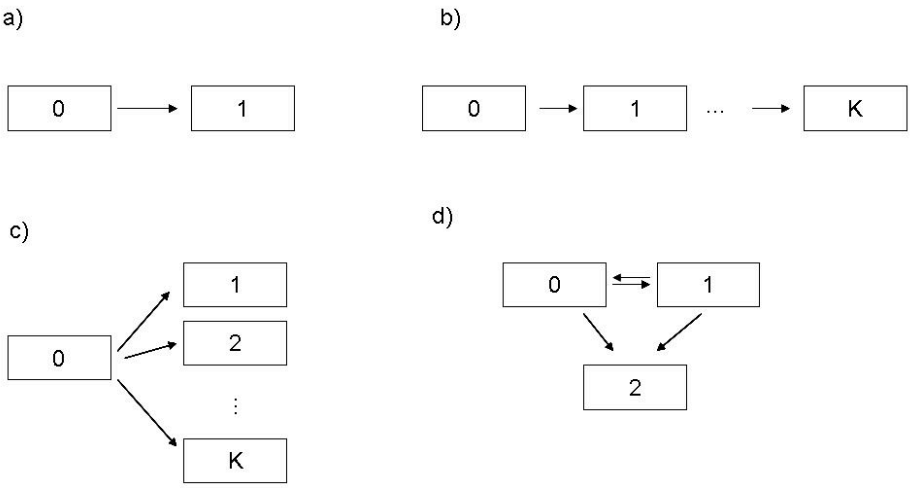

Figure 1.1: Examples of multi-state diagrams: (a)failure time; (b)progressive; (c)competing risk; (d)illness-death.

employed" and "out of the labour force". It may be of interest to examine the probability distributions of the time to leave the "unemployed" state to the other two remaining states separately, that is, the duration of an unemployment spell before transitioning to being employed or to out of the labour force. It may also be of interest to analyze its relation with covariates such as education level and marital status.

The diagrams in Figure 1.1 illustrate some types of multistate processes. The simplest process, (a), corresponds to the failure time model discussed in the survival analysis section; (b) illustrates a progressive model where states occur in an ordered sequence, useful for representing a sequence of events; (c) is the competing risks or multiple failure mode model, consisting of $K-1$ absorbing states; and finally (d), which in biostatisics is often called an illness-death model since it illustrates the states under study indicated by, for instance, 0-healthy, 1-sick, 2-deceased (absorbing). The first example mentioned above can be represented by the diagram in (d) and the second example may be represented in the competing risks setting in (c).

From the preceding paragraphs, it becomes natural to see that multistate models are addressed through a random variable $Y_i(t)$ that indicates the state in the set $\{1, 2, ..., K\}$ which the individual i is in at time t. An equivalent way to keep track

of this kind of process is through event occurrence counting processes, which instead uses $N_{ij}(t)$ to record the number of times an individual i has an event of type j, $j = 1, 2, ..., J$ at time t. Here, the events are the J different types of transitions that can be made between states.

Although they are mathematically equivalent, the multistate and event occurrence approaches are used for different objectives. The former is commonly used when interest lies in studying duration times. Note that from $Y_i(t)$ it is possible to obtain the length of sojourn in a specific state, that is, the elapsed time for the process to leave a state and make a transition to another. The event occurrence approach is used when the number of visits to the states is rather in question. In this work, the focus will be on the study of duration times in the multistate framework.

Consider the external covariates $x_i(t)$ and let the history of the process up to time t be denoted by $H_i(t) = \{Y_i(s), 0 \leq s < t\}$. If it is assumed that two events cannot occur simultaneously, then the full event history process can be described, from the multistate persective, by the transition intensity functions defined as:

$$\lambda_{ikl}(t|x_i(t), H_i(t)) = \lim_{\Delta t \to 0} \frac{Pr\left\{Y_i(t + \Delta t) = l | Y_i(t^-) = k, x_i(t), H_i(t)\right\}}{\Delta t} \qquad (1.13)$$

where $k \neq l$, are valued on the state space $\{1, 2, ..., K\}$.

Simplifications of the intensity function are often used in practice, such as Markov models where it only depends on $x(t)$ or t and Semi-Markov (renewal) models where the intensity depends on $H_i(t)$ only through the elapsed time since the most recent transition and on $x(t)$.

The survival model described in section 1.2 is a special case of multistate models, since it can be considered as a transitional model with two states (see Figure 1.1 (a)), where the only possible transition is from state 1 to state 2. Defining T_i as the duration of the i individual's visit to state 1, the intensity function in (1.13) is simplified to $\lambda_i(t|x_i(t)) = \lim_{\Delta t \to 0} Pr\left\{T_i < t + \Delta t | T_i \geq t, x_i(t)\right\} / \Delta t$, which is the definition of the hazard function $h(t)$ used in (1.1).

The competing risks setting (Figure 1.1, (c)) is modelled as follows. Suppose

that a continuous failure time T is subject to several modes or causes of failure CF, $CF \in \{1, 2, ..., K\}$ and also to right censoring (recall the "unemployed" vs. "employed" and "out of labour force" example). One way of fully specifying the distribution of (T, CF) is by the mode-specific hazard function

$$\lambda_j(t) = \lim_{\Delta t \to 0} \frac{Pr(T < t + \Delta t, CF = j | T \geq t)}{\Delta t};\ j = 1, 2, ..., K. \tag{1.14}$$

The marginal hazard function for T is $\lambda(t) = \sum_{j=1}^{K} \lambda_j(t)$ and the marginal survivor function is $S(t) = \exp(-\Lambda(t))$, where $\Lambda(t) = \sum_{j=1}^{K} \Lambda_j(t)$ is the cumulative hazard function for T. The marginal survivor function can also be expressed as $S(t) = \prod_{j=1}^{k} G_j(t)$, where $G_j(t) = \exp(-\Lambda_j(t))$. The function $G_j(t)$ is not a survivor function for any observable random variable; however, it is used for convenience as described below, in expression (1.16).

The mode-specific distribution and density functions, referred to as the subdistribution and subdensity functions are usually of interest for analyzing a specific mode of failure. They are given by, respectively:

$$F_j(t) = \Pr(T \leq t, CF = j) = \int_0^t \lambda_j(u) S(u) du, \text{ and } f_j(t) = F_j'(t) = \lambda_j(t) S(t).$$

Information is collected from a random sample of size n that gives either $(T_i = t_i, CF_i)$ or $T_i > t_i$. Defining $\delta_i = 1$ if t_i is a failure time and 0 if t_i is a censoring time, and assuming independent censoring, then the likelihood function is given by:

$$L = \prod_{i=1}^{n} f_{CF_i}(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \tag{1.15}$$

$$= \prod_{j=1}^{k} \prod_{i=1}^{n} g_j(t_i)^{\delta_{ij}} G_j(t_i)^{1-\delta_{ij}} = \prod_{j=1}^{k} L_j \tag{1.16}$$

where $\delta_{ij} = I(CF_i = j, \delta_i = 1)$, $f_j(t) = \lambda_j(t) S(t)$ and $g_j(t) = \lambda_j(t) G_j(t) = -G_j'(t)$. As indicated earlier, the functions $g_j(t)$ and $G_j(t)$ in (1.16) do not correspond to any observable random variable; however, the obtained likelihood expression has the standard form for a survival time distribution. If we assume that these functions involve separate parameters θ_j for $j = 1, 2, ..., k$, then usual procedures can be used and estimation can be done for each type of failure separately through L_j. In this

case, a failure of type j at time t_i is treated as a failure for estimation of θ_j and a failure of any other mode at t_i is treated as censoring.

Parametric or semi-parametric models for the intensity functions in (1.13), and (1.14) can be specified. In the parametric case, asymptotic properties for the score function and for the maximum likelihood estimate can be used for inference. The semi-parametric multiplicative formulation of the intensity function can be analyzed using partial likelihood arguments described earlier. For a more detailed discussion on estimation in multistate models, see Kalbfleisch and Prentice [25], Blossfeld et al. [7] and Lawless [32].

1.4 Inference from survey data

Survey data are usually collected based on multi-stage stratification and cluster sampling designs as well as unequal probabilities of selection. To account for this, standard design-based sampling theory for survey data involves the use of design weights in the estimation of descriptive quantities such as population means and proportions. In contrast, the analytical study of survey data where interest lies in examining the relation of other variables with a response, leads to the use of more complex, model-based estimation techniques. There is controversy on whether survey weights are necessary in the use of model-based estimation techniques (for example, see Korn and Graubard, [28], Gelman ([22], Little [41], [42]).

Analytic inference can be performed from the superpopulation or finite population perspectives. The former regards the survey population as a random sample from an infinite universe and this randomness is accounted for in statistical inference. The latter treats the population as finite; the data obtained are considered fixed and the only realization of a random variable is through the sampling mechanism, which determines the probability an individual is included in the sample.

The superpopulation based methods described in the following section have to do with the idea of ignorable sampling. In line with Chambers and Skinner

([12], p.7), suppose that we have a population of size N and the sample inclusion indicator vector $I_U = (I_1, ..., I_N)$, $I_i = I(i \in S)$. Also, let $Z_U = (Z_1, ..., Z_N)$ denote the matrix of design related factors Z_i such as cluster and stratum information; and $Y_U = (Y_1, ..., Y_N)$ the matrix of response vectors Y_i for each individual. Suppose that the realizations of the random variables (I_U, Y_U, Z_U) are denoted by (i_U, y_U, z_U). The joint distribution indexed by the parameters (ϕ, ψ) is:

$$f(i_U|z_U, y_U)f(y_U|z_U; \phi)f(z_U; \psi).$$

Further, let Y_{obs} denote the responses from the sampled individuals and Y_{miss} those from the individuals that were not included in the sample. The observed data are given by $(i_U, y_{\text{obs}}, z_U)$ and the likelihood for (ϕ, ψ) is:

$$L(\phi, \psi) \propto \int f(i_U|z_U, y_U)f(y_U|z_U; \phi)f(z_U; \psi)dy_{\text{miss}}. \tag{1.17}$$

For the sampling design to be ignorable, it is necessary that $f(i_U|z_U, y_U) = f(i_U|z_U)$ and that the function $f(i_U|z_U)$ does not depend on the parameters (ϕ, ψ). If these conditions are met, then it is possible to make likelihood-based inference treating i_U as fixed and therefore discarding $f(i_U|z_U, y_U)$ from (1.17). In more general cases, additional covariates X_U can be introduced so the model for responses is $f(y_U|z_U, x_U; \phi)$ and the requirement for design ignorability is $f(i_U|z_U, x_U, y_U) = f(i_U|z_U)$. Further discussion about ignorable designs can be found in Pfefferman [46] and Binder and Roberts [5].

As will be discussed in more detail in the next section, the finite population perspective for analytic inference is based on the notion of estimating a population quantity that is an implicit function of the parameter of interest and is referred to in sampling as pseudo-likelihood inference.

1.5 Surveys and survival analysis

As mentioned earlier, survival analysis methodology can be applied in the case of single durations per individual, and has been described for the non-survey setting

in section 1.2 . Extensions to survey data may proceed from the superpopulation or the finite population perspectives. In particular, this section provides a summary of survey methods based on the Cox PH formulation, under the assumption of independent loss to follow-up. Contributions include those of Binder [4] in the finite population framework, a combination of finite and superpopulation based inference given in Lin [38] and superpopulation based inference given in Boudreau and Lawless [9] and Lawless [32].

With regard to the analysis of multiple durations, the contributions in the literature have been found to be sparse. Some of these include Blossfeld and Hamerle [6], Hamerle [23] and Kovacevic and Roberts [29] regarding the marginal estimation via the Cox model of unemployment duration distributions. The analysis of successive duration times will be discussed in more detail in Chapter 2.

Superpopulation inference

Suppose that a finite population $\mathcal{U} = \{1, ..., N\}$ is divided into R disjoint strata $\{\mathcal{U}_1, ..., \mathcal{U}_R\}$ and that the primary sampling units (PSU's) are clusters of individuals that are selected within the strata, according to a given sampling design. Then, subsamples S_{kr} are chosen within cluster k in stratum r, $k = 1, ..., K_r$ and $r = 1, ..., R$. The sample S is then expressed by:

$$S = \bigcup_{r=1}^{R} \bigcup_{k=1}^{K_r} S_{kr}.$$

Let $t_i = \min(T_i, C_i)$ represent either the time to an event (T_i) or the censoring time (C_i), $\delta_i = I(T_i \leq C_i)$ indicate status ($\delta_i = 0$ if censoring is present) and x_i represent the covariates of individual i. Therefore, the information collected from an individual that experienced the event of interest in the sample consists of $\{t_i, \delta_i, x_i\}$. Let θ be the parameter of interest and $\lambda(t_i|x_i;\theta)$, $f(t_i|x_i;\theta)$ and $S(t_i|x_i;\theta)$ denote the hazard, density and survivor functions given x_i, respectively. These models represent marginal distributions for an individual. Given the considerable heterogeneous structure in the population, the model may include stratum information

among the covariates x_i, and may apply only to a certain subgroup of the population. Moreover, the T_i are not in general independent, given x_i. In this thesis, we do not attempt to model association between individuals, but allow for the association in inferences about θ.

Suppose that the model of interest gives rise to an estimating function for each individual i, denoted by $U_i(\theta)$. Estimation from the superpopulation perspective is possible through (Lawless,[31], p.234):

$$U(\theta) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i \in S_{kr}} U_i(\theta) = 0 \tag{1.18}$$

where S_{kr} denotes the sample of individuals in the kth stratum and rth cluster, defined earlier. It is assumed that any stratum effects are modelled via the covariates x_i, so that $E\big(U_i(\theta)\big) = 0$ for i in each S_{kr}.

The estimating equation in (1.18) is unbiased under the assumption of correctness of the model, ignorable sampling (described in the preceding section) and ignorable censoring, that is, when $Pr(T_i|C_i, I_i = 1, x_i,) = Pr(T_i|I_i = 1, x_i)$. The estimator $\hat{\theta}$ is asymptotically normal with variance estimated by:

$$\hat{V}(\hat{\theta}) = I(\hat{\theta})^{-1} \hat{V}\big(U(\theta)\big) I(\hat{\theta})^{-1}, \tag{1.19}$$

where $I(\theta) = -\partial U(\theta)/\partial\theta$ is the observed information matrix and

$$\hat{V}\big(U(\theta)\big) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} \Big(\sum_{i \in S_{kr}} U_i(\hat{\theta}) \Big) \Big(\sum_{i \in S_{kr}} U_i(\hat{\theta}) \Big)'. \tag{1.20}$$

The Cox PH model can also be handled via an estimating function like (1.18). For this case, we will illustrate the approach for a stratified Cox model. For a set of strata $r = 1, 2, ..., R^0$, suppose the hazard function is then $\lambda_{ir}(t|x_i(t)) = \lambda_{0r}(t)\exp(x'_{ir}\beta)$ referring to the ith person in stratum r, where β is a $p \times 1$ vector of regression parameters and the $\lambda_{0r}(t)$ are arbitrary baseline hazard functions. It is important to note that the strata specified in this model usually do not represent the lowest stratum level in the survey design.

Define the risk indicator as $Y_i(t) = I(t_i \geq t)$. Then the estimating function for β is (Boudreau and Lawless [9]):

$$U(\beta) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i \in S_{kr}} \delta_i \left(x_i(t_i) - \frac{S_r^{(1)}(t_i, \beta)}{S_r^{(0)}(t_i, \beta)} \right), \quad \text{where} \tag{1.21}$$

$$S_r^{(0)}(t, \beta) = \sum_{k=1}^{K_r} \sum_{i \in S_{kr}} Y_i(t) \exp(x_i(t)'\beta) \quad \text{and} \tag{1.22}$$

$$S_r^{(1)}(t, \beta) = \sum_{k=1}^{K_r} \sum_{i \in S_{kr}} Y_i(t) x_i(t) \exp(x_i'\beta). \tag{1.23}$$

Left truncation frequently arises with duration data from surveys. This occurs when the spell for which a duration or failure time is defined started at time l_i prior to the observation period; in that case we know that $T_i \geq l_i$ and we condition on this fact. For example, an individual may have been jobless for a time l_i prior to the time at which they join a study. Likelihoods like (1.1), (1.6) and (1.15) can be adjusted for left truncation. In (1.21), left truncation is readily introduced by just redefining the risk indicator as $Y_i(t) = I(l_i \leq t \leq t_i)$. Truncation times are assumed to be independent of the event times, given covariates x_i.

Even though the estimating equation $U(\beta) = 0$ was originally designed for within and between cluster independence of $\{T_i,\ i \in S_{kr}\}$ for $\hat{\beta}$ to be consistent, Boudreau and Lawless [9] show that in the case of within cluster association, $U(\beta) = 0$ is asymptotically unbiased and so can be used for consistent estimation of β. This applies for a large number of clusters and bounded cluster size.

Similarly, it is shown in Boudreau and Lawless [9] that Breslow-Aalen estimates of the baseline cumulative hazard functions $\Lambda_{0r}(t)$ are consistent for arbitrarily large number of clusters with a bounded size within each stratum, and have the form:

$$\hat{\Lambda}_{0r}(t, \hat{\beta}) = \sum_{k=1}^{K_r} \sum_{i \in S_{kr}} \frac{\delta_i I(t_i \leq t)}{S_r^{(0)}(t_i, \hat{\beta})} \tag{1.24}$$

Point estimators of β and $\Lambda_{0r}(t)$ as well as variance estimates can be obtained from standard survival analysis software. The flexibility of the coxph function in

SPlus/R can be used in conjunction with the strata and cluster options to perform these procedures.

A weighted version of $U(\beta)$ can be used when sampling is non-ignorable, so that the estimating function (1.21) is not asymptotically unbiased. The weight w_i is proportional to the inverse of the sample inclusion probability $\pi_i = P(i \in S)$ and the weighted versions of (1.21), (1.22), (1.23) can be used to give an estimate $\hat{\beta}_W$ and the respective weighted version of (1.24) to give $\hat{\Lambda}_{0r_W}(t, \hat{\beta}_W)$. Asymptotic variance estimators are obtained under the same line of development as for the aymptotic variance of the unweighted estimates from (1.21) and (1.24).

Finite population inference

For many analyses of multivariate survey data, it is convenient to define the parameters of interest as implicit functions of population totals, rather than explicitly, as is done usually for descriptive inference (Binder [3]). As before, suppose that a size N population is divided into $r = 1, ..., R$ strata of size N_r and that stratum r in the population is composed of $\mathcal{K}_r$ clusters. Also, let $\mathcal{U}_{kr}$ denote the subpopulation corresponding to the kth cluster in the rth stratum. Let θ denote the parameter of interest to be defined implicitly through the following population quantity:

$$U_{\mathcal{U}}(\theta) = \sum_{r=1}^{R} \sum_{k=1}^{\mathcal{K}_r} \sum_{i \in \mathcal{U}_{kr}} U_i(\theta), \tag{1.25}$$

where $U_i(\theta)$ is the pseudo-score contribution from individual i. It is the finite population version of the score function in (1.18), and as such, does not have the same interpretation provided by the superpopulation framework, in the sense of being a function of random quantities. Let θ_N denote the solution to $U_{\mathcal{U}}(\theta) = 0$.

Suppose that K_r clusters are selected from the $\mathcal{K}_r$ clusters in stratum r of the population, for $r = 1, ..., R$. The sample estimate of the population quantity in (1.25) is given by

$$\hat{U}_{\mathcal{U}}(\theta) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i \in S_{kr}} \mathrm{w}_i U_i(\theta), \tag{1.26}$$

where the w_i are sampling weights corresponding to the sample inclusion probabilities. The pseudo-maximum likelihood estimate $\hat{\theta}$ satisfies $\hat{U}_{\mathcal{U}}(\hat{\theta}_N) = 0$. Pseudo-likelihood inference can be approached from design-based or model-based perspectives. Variance estimation is usually performed using a combination of a Taylor series linearisation argument and an appropriate method (design or model based) for estimating the variance of $\hat{U}_{\mathcal{U}}(\theta)$. The "sandwich" variance estimator of $\hat{\theta}$ that is obtained from the linearisation has the form:

$$\hat{V}(\hat{\theta}) = \left\{ \left(\frac{\partial \hat{U}_{\mathcal{U}}(\theta)}{\partial \theta} \right)^{-1} \hat{V}\left(\hat{U}_{\mathcal{U}}(\theta) - U_{\mathcal{U}}(\theta) \right) \left(\frac{\partial \hat{U}_{\mathcal{U}}(\theta)}{\partial \theta} \right)^{-1} \right\} \Bigg|_{\theta=\hat{\theta}_N}, \qquad (1.27)$$

where $\hat{\theta}$ is the solution to $\hat{U}_{\mathcal{U}}(\theta) = 0$ and $\hat{V}[\hat{U}_{\mathcal{U}}(\theta) - U_{\mathcal{U}}(\theta)]$ is a corresponding estimator of the variance of $\hat{U}_{\mathcal{U}}(\theta) - U_{\mathcal{U}}(\theta)$ (Binder, [3]; Chambers [11]).

Binder [3] explains differences between implicit and explicit parameters and provides a discussion of the pseudo-likelihood theory as well as variance estimation methods of the form in (1.27) for generalized linear models. The case of Cox's PH partial likelihood is as follows.

The pseudo-likelihood theory described above gives the following expressions for the Cox proportional hazards model (Binder,[4]). The pseudo-score function for the population is given by:

$$U_{\mathcal{U}}(B) = \sum_{r=1}^{R} \sum_{k=1}^{\mathcal{K}_r} \sum_{i \in \mathcal{U}_{rk}} \delta_i \left(x_i - \frac{S^{(1)}(t_i, B)}{S^{(0)}(t_i, B)} \right), \quad \text{where} \qquad (1.28)$$

$$S^{(0)}(t, B) = \sum_{r=1}^{R} \sum_{k=1}^{\mathcal{K}_r} \sum_{i \in \mathcal{U}_{rk}} Y_i(t) \exp(x_i' B),$$

$$S^{(1)}(t, B) = \sum_{r=1}^{R} \sum_{k=1}^{\mathcal{K}_r} \sum_{i \in \mathcal{U}_{rk}} Y_i(t) x_i \exp(x_i' B),$$

and where t_i is the failure or censoring time, δ_i is the censoring indicator, $Y_i(t) = I(t_i \geq t)$ is the risk indicator and x_i are the covariates of individual i. Note that these are the finite population version of score functions in (1.10) and B is the finite population parameter of interest, that is, the solution to $U_{\mathcal{U}}(B) = 0$.

The estimating equation that results from the sample estimate of (1.28), with sampling weights w_i which are constructed so that the weighted sums are approximately unbiased and consistent estimates of the corresponding means over the finite population and giving the pseudo-maximum likelihood estimate $\hat{B}$, has the form:

$$\hat{U}_{\mathcal{U}}(\hat{B}) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i \in S_{kr}} \mathrm{w}_i \delta_i \left(x_i - \frac{\hat{S}^{(1)}(t_i, \hat{B})}{\hat{S}^{(0)}(t_i, \hat{B})} \right) = 0, \qquad \text{where} \tag{1.29}$$

$$\hat{S}^{(0)}(t, \hat{B}) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i \in S_{kr}} \mathrm{w}_i Y_i(t) \exp(x_i' \hat{B}),$$

$$\hat{S}^{(1)}(t_i, \hat{B}) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i \in S_{kr}} \mathrm{w}_i Y_i(t) x_i \exp(x_i' \hat{B}).$$

Note that the above expression has the form:

$$\hat{U}_{\mathcal{U}}(\hat{B}) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i \in S_{kr}} \mathrm{w}_i U_i(\hat{B}, \mathrm{w}_i) = 0. \tag{1.30}$$

Binder uses (1.30) to approximate (1.26) in the sandwich estimator defined in (1.27) with a design-based estimate of $V\big(\hat{U}_{\mathcal{U}}(B)\big)$.

It is important to note that the information from the population $\{t_i, \delta_i, x_i\}$ $(i = 1, 2, ..., N)$ is considered as fixed in this context, and that the parameter value B does not have an exact hazard ratio interpretation, as possessed by the regression parameter β in (1.8). We note that, in fact, the censoring processes often apply only to individuals in the sample, for example, when they result from loss to follow-up. Furthermore, the C_i are random at the time the sample is drawn. Thus, the idea of fixed functions $U_i(\theta)$, $i = 1, \ldots, N$ is not completely true here, since the line between finite and superpopulation inference is not clear.

Lin [38] extends Binder's variance estimate to the superpopulation approach and provides a formal justification of the proposed variance. This inference procedure treats the information of the survey population as a random sample $\{t_i, \delta_i, x_i\}$ $(i = 1, 2, ..., N)$ and is therefore conceived from the superpopulation perspective. This allows for a straightforward interpretation of the estimated covariate effects.

Lin's variance consists of Binder's variance plus one more term that accounts for randomness coming from the superpopulation. Boudreau and Lawless [9] have found that in settings where the population size is large with respect to the sample size, their variance estimates do not significantly differ from Lin's, and neither differ much from Binder's.

Both superpopulation and finite-population methods can be carried out in the SUDAAN package, via the `SURVIVAL` procedure. Since SUDAAN is a package designed for survey data, it is possible to specify sampling design features in the point and variance estimation from Cox's porportional hazards models. SAS provides Cox's regression and Binder's variance estimates through the `cov` option in the `PHREG` procedure allowing for the use of case-specific weights. Similarly, SPlus and R can be used via the `coxph` function with the `cluster` and `strata` options, where weights are also allowed.

In cases where within-cluster dependence does not affect the estimation results dramatically (as in Boudreau [8]), standard diagnostic and model checking methods can be used. For cases like this, residuals and diagnostic plots to assess goodness of fit, methods to identify influential observations and to examine the functional form and the proportional hazards assumption can be be carried out from both the weighted and unweighted methods described here. For a detailed summary on model checks and diagnostics, see Therneau and Grambsch ([57], chapters 4-7), where SAS and SPlus functions are used.

1.6 Longitudinal surveys

Longitudinal surveys add a new dimension to cross sectional surveys in that they provide information about the evolution of a population over time. The latter are useful in describing a population in a specific moment, and in this sense we can say that they provide a static snapshot of the population, while longitudinal surveys aim to portray the dynamic nature of lifetime processes for individuals in

the population.

One example of longitudinal surveys is the Survey of Labour and Income Dynamics (SLID) of Statistics Canada. It provides information about transitions in jobs, income and family events experienced by Canadian individuals. Samples from SLID are selected from the Labour Force Survey (LFS) and thus share the latter's complex sample design. Individuals are interviewed annually for a period of six years, and the information is gathered retrospectively regarding events that occurred during the year that elapsed since the last interview. Information about the LFS and SLID can be found online at www.statcan.gc.ca/imdb-bmdi/3701-eng.htm and www.statcan.gc.ca/pub/75f0011x/4060256-eng.htm, respectively. A comprehensive summary of SLID can be found in Boudreau, [**?**].

Another example is the Millennium Cohort Study (MCS), providing data from children growing up in the four countries of the United Kingdom. This is a complex survey that aims at understanding the social and economic conditions surrounding birth and early childhood and collects information regarding the develpment of children that were born in 2000 and 2001. Information has so far been collected for children at ages 9 months, 3, 5 and 7 years old, in the years 2001/2, 2004/5, 2006, and 2008, respectively. The fifth wave is scheduled to take place in 2012, when the cohort children will be age 11. Online information about the MCS can be found at www.cls.ioe.ac.uk/text.asp?section=000100020001.

Data from longitudinal surveys are often collected in spaced interviews over a long period of time. This may lead to partial or total loss of information from individuals over the observation period. As a result, not only is loss to follow-up common by the end of the survey, but also it may be dependent on events or covariates of interest. In the above examples, loss to follow-up becomes substantial over time. SLID samples are typically in the 25-30 percent range of loss to follow-up by the end of the six years. In the MCS study, there was a loss to follow-up rate of 28 percent in the first wave and 42 percent by the second wave (Plewis,[47]). Dependent loss to follow-up has been considered by many authors in the context of

continuous and binary outcomes (e.g. Robins et al., [51]; Miller et al., [44]; Preisser et al., [48]); however, duration analysis where data are collected retrospectively has not been considered.

The observational framework for the analysis of durations can be described as follows. Individuals selected for a panel are seen at times $t = 0, 1, 2, \ldots, M$ over a period $(0, M]$. In SLID for example, $M = 6$ years and $t = 1, \ldots 6$ represent the years of a panel, 1996-2001, 1999-2004, etc. At time t, information about the event of interest and covariates $D_i(t)$ is collected retrospectively from the period $(t-1, t]$. At the initial visit $(t = 0)$, baseline information is collected which may or may not include details of events that started before $t = 0$. Individuals are subject to be missing from the survey at any interview time $t \leq M$. In some cases, the missing data pattern may be monotone, this means that individuals do not return to the survey after being absent once. In some other cases, it is possible to allow individuals to return to the study. The methods that are presented here will focus only in the former patter of missing data. In SLID for example, individuals are allowed to return after no more than two consecutive absent interviews. In our analyses from SLID however, only the information of individuals up to the first time they missed an interview or labour information is not given is considered. Methods for non-monotone patterns of missing data are of practical interest and techniques need to be developed, this will be further discussed in chapter 8.

When modelling durations, loss to follow-up should be considered if it is suspected that it is related to the event history and covariates. Since a duration typically overlaps more than one interval $(t-1, t]$, we require a weighted analysis with time-varying weights. In this observational framework, loss to follow-up can be modelled for each interview time by the probability that an individual is observed at time t, given that he or she was also observed at time $t-1$. Covariate information up to time $t-1$ in the model for loss to follow-up at time t may also include event related and sampling design variables. As will be discussed in more detail in Chapters 3 and 4, it is important that this covariate information is

enough so that it is reasonable to assume that the event of interest (duration) is conditionally independent of loss to follow-up.

1.7 Outline of the thesis

Chapter 2 provides a discussion about duration analysis which is based on the survival analysis and multistate models theory given in sections 1.2 and 1.3. It gives the basics of conditional estimation in classical settings and a discussion about issues involving marginal analysis of sequences of durations from survey data. The issue of dependent loss to follow-up and the importance of accounting for it when performing analysis of durations is discussed, as well as an introduction of the Inverse Probability of Censoring Weighted (IPCW) method.

The main objective of Chapter 3 is to describe some of the proposed methods for analysis in the context of duration and event history analysis, without considering sampling theory. The observational framework within which the methods will be applied is described and a section is dedicated to describe the model that is employed for estimation of the IPC weights. Examples are presented regarding the application of the IPCW method to the analysis of duration time distributions.

The contents of Chapter 4 are the basis for the methods that are proposed in Chapters 5 and 6, where Kaplan-Meier and Cox PH models are discussed. It gives the IPCW techniques extended to the context of duration variables, which are based on estimating function theory for parametric models. A simulation study illustrates the performance of the IPCW under dependent loss to follow-up, regarding log-Normal durations. The results show that the use of IPC weights reduces bias in the estimation of regression coefficients and that the proposed variance estimation method performs well. Finally, an extension of the methodology to survey data is provided.

Chapter 5 describes the methods for applying the Kaplan-Meier estimator in the context of survey data. The methods for variance estimation are based on

the parametric modelling presented in Chapter 4. A simulation study is presented to assess IPCW estimates of the survivor function based on multiple spells from individuals sampled from a finite population. This simulation shows our methods give good results in the presence of non-ignorable sampling design (in the lines of section 1.4) and dependent loss to follow-up.

Variance estimation of estimates from the Cox PH model is discussed in Chapter 6. The parametric methods described in Chapter 4 can be used when using a Piecewise Constant (PC) model as an approximation to the Cox PH model. Estimating functions for both the Cox PH and the PC models are described as well as the proposed variance estimation procedure. A simulation study is presented where the PC approximation to the Cox PH model is assessed, with good overall results, indicating the feasibility of the application of this method on real data sets.

Implementation of the methods for Kaplan-Meier estimates and for Cox PH regression coefficients discussed in chapters 5 and 6 is carried out for jobless spells from the Survey of Labour and Income Dynamics (SLID). This chapter gives a broad discussion regarding the issues that have been encountered while analyzing SLID data, which are common in longitudinal surveys. This chapter also gives a descriptive analysis of the SLID data regarding loss to follow-up, missing data and characteristics of the jobless spells, focusing on members of panel 3, which covers the period from 1999 to 2004. Implementation of the proposed methodology is performed on jobless spells from individuals residing in Ontario and Quebec in the year 1999.

Final remarks and conclusions are presented in Chapter 8. This chapter gives pointers for future research, for example, for the generalization of the monotone missing data pattern that was assumed in the proposed methods, as well as procedures for handling missing data in covariates and response variables in the context of duration analysis of survey data.

Chapter 2

Analysis of Durations from Longitudinal Survey Data

2.1 Motivation

One of the main objectives of this work is to propose methods for the analysis of relevant distributions for single durations and also for the analysis and modelling of sequences of duration times, or multiple durations. A sequence of durations may occur as successive state duration times for a sequence of states. For example, a heart transplant study where durations from a sequence of states given by "admitted to program, pre-transplant", "alive, post-transplant", "dead" are observed. Another example is given by the sequence of durations of the sojourns in the states "disease-free" and "recurrence" in a cancer study. There are also types of sequences where the successive durations are observed in a specific state. For instance, the sequence of durations of unemployment spells, the durations of maternity leaves in working women, the duration of quitting attempts in smokers.

The standard survival methods for single durations discussed in the preceding chapter consitute the building blocks for the analysis of sequences of durations. Sometimes it is reasonable to assume that the within-individual duration times are independent. Independence is attained if the time of occurrence of an event is un-

affected by the times of the preceding events, that is, by previous event history. However, it usually is the case that multiple observations in one individual are subject to a possible interdependence. For example, the duration of an unemployment spell may be related to the length and the number of previous unemployment episodes an individual has experienced. Consequently, the analysis of sequences of durations may not only include an examination of the effects of fixed or time-varying covariates and association within clusters of individuals, but may include also within-individual dependence.

Section 2.2 gives the basics of conditional estimation in classical settings. As motivation, suppose that it is reasonable to fit a model for the first jobless spells for individuals in a given year. When considering any subsequent spells however, covariate information such as the start time of the spells, and times and durations of preceding spells should be considered in the model. A limitation of this method is that after conditioning on previous events, generalizations of the results to address average population features become difficult. However, this approach is essential if we wish to understand the dynamics of employment on an individual level.

Survey studies often focus on answering questions regarding characteristics of the population. For instance, SLID data may be useful in estimating the distribution of unemployment spells that begin in a specific calendar year or the joint probability distribution of employment and unemployment spells that begin in a specific calendar year given a set of covariates. Such population level features are easier to address than marginal features associated with specific individuals. Section 2.3 gives a discussion about issues involving marginal analysis of sequences of durations from survey data.

2.2 Conditional analysis of sequences of duration times

Let $(0, M]$ denote the potential follow-up period for individuals $i = 1, 2, ..., n$ and suppose m_i duration times Y_{ij} $(j = 1, 2, ..., m_i)$ occur for individual i. Let u_{ij} and v_{ij} denote the start and end times of the jth spell's duration, such that $u_{i1} < v_{i1} \leq u_{i2} < v_{i2} \leq u_{i3}$, etc. The duration times are then $Y_{ij} = v_{ij} - u_{ij}$ and the time the person was last seen is denoted by C_i. The observed duration time is then $y_{ij} = \min(Y_{ij}, C_i - u_{ij})$ and the non-censored indicator is given by $\delta_{ij} = I(Y_{ij} \leq C_i - u_{ij})$. Note that only the last duration can be censored in this framework.

Assume that the within-individual duration times have conditional distributions:

$$F_j(y|z_{ij}) = Pr(Y_{ij} \leq y|z_{ij}) \qquad j = 1, 2, ..., m_i \tag{2.1}$$

where the vector z_{ij} may consist of covariates x_{ij} and features of previous spells and previous duration times $y_i^{(j-1)} = \{y_{i1}, y_{i2}, ..., y_{i,j-1}\}$. For example, z_{ij} may include not only $y_i^{(j-1)}$, but also the start and end times of previous spells. Since the process to which the durations Y_{ij} belong involves a set of two or more states, then the vector z_{ij} could also include information regarding the sojourn in other states. For example, suppose that we were analyzing a process of two states, "employed" and "unemployed" and that our interest lies in the durations of the visits spent in the "unemployed" state. Then the vector z_{ij} could include not only the previous durations in the state of interest, but also the durations and the start times in the "employed" state.

Moreover, let $f_j(y_{ij}|z_{ij})$, $h_j(y_{ij}|z_{ij})$, and $S_j(y_{ij}|z_{ij})$ represent the jth duration time density, hazard and survivor function given the vector z_{ij}, respectively. Assuming that Y_{ij} is conditionally independent of $Y_i^{(j-1)}$ given z_{ij} and that the last duration time is possibly right censored, we can express the overall likelihood from

n independent individuals by:

$$L = \prod_{i=1}^{n} \prod_{j=1}^{m_i-1} f_j(y_{ij}|z_{ij}) S_j(y_{im_i}|z_{ij}), \tag{2.2}$$

$$= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{m_i-1} h_j(y_{ij}|z_{ij}) \exp(-H_j(y_{ij}|z_{ij})) \right\} \exp(-H_j(y_{im_i}|z_{ij})). \tag{2.3}$$

Note that for every observed spell y_{ij}, the vector z_{ij} may include the start time of the spell, u_{ij}. For the first observed spell y_{i1} in $(0, M]$, we assume that the start time u_{i1} is known, even if $u_{i1} < 0$.

Standard survival analysis methods and software on Accelerated Failure Time (AFT) and Proportional Hazards (PH) regression models can be applied using the likelihood function for right censored data from expressions (2.2) and (2.3). The AFT models described in the previous chapter are easily used by letting $Y_{ij}^* = \log(Y_{ij})$ and defining the independent and identically distributed random variables ϵ_{ij} for $i = 1, 2, ..., n$ to give:

$$Y_{ij}^* = \beta_{0j} + z_{ij}'\beta_j + \sigma_j \epsilon_{ij} \qquad j = 1, 2, ..., m_i. \tag{2.4}$$

Just as described earlier, the error distribution is usually taken to be standard normal, extreme value or logistic.

The partial likelihood for the semi-parametric multiplicative model

$$h_{ij}(y|z_{ij}) = h_{0j}(y) \exp(z_{ij}'\beta_j), \tag{2.5}$$

is similar to the partial likelihood in (1.9). Note that here covariates are assumed fixed across a spell; however, the methods discussed can be extended to time varying covariates. The partial likelihood for estimating β_j is:

$$L_j(\beta_j) = \prod_{i=1}^{n} \left\{ \frac{\exp(z_{ij}'\beta_j)}{\sum_{l=1}^{n} \delta_{lj} I(y_{lj} \geq y_{ij}) \exp(z_{lj}'\beta_j)} \right\}^{\delta_{i,j+1}}, \tag{2.6}$$

where $\delta_{ij} = 1$ if the $(j-1)$st event was observed from individual i and $\delta_{ij} = 0$ otherwise. A similar idea applies for the estimator of the baseline cumulative hazard functions. Using $\hat{\beta}_j$ from maximizing (2.6) we get:

$$\hat{H}_{0j}(y) = \sum_{i=1}^{n} \left\{ \frac{\delta_{i,j+1} I(y_{ij} \leq y)}{\sum_{l=1}^{n} \delta_{lj} I(y_{lj} \geq y_{ij}) \exp(z_{lj}'\hat{\beta}_j)} \right\}. \tag{2.7}$$

Interest may reside in estimating quantities based on the above expression, such as the survivor function:

$$\hat{S}_j(y|z_{ij}) = \exp\left\{-\hat{H}_{0j}(y)e^{z'_{ij}\hat{\beta}}\right\}. \tag{2.8}$$

The preceding discussion assumes that the same family of models applies to the jth duration time of any individual with one or more durations. This may not always be sensible, because two individuals may have had different event histories prior to time $t = 0$, and different histories over $(0, M]$. The specification of models for sequences of durations is dependent on the setting and on the objectives of analysis, and it is difficult to give a general treatment.

Sometimes the process has started before the observation period $(0, M]$, that is, $u_{i1} \leq 0$. Earlier we indicated that Y_{i1} may be subject to left truncation, and described how to handle this for the Cox model. In the case of parametric models, the following procedure is applied. Since the first duration time is defined as $Y_{i1} = v_{i1} - u_{i1}$, then $Y_{i1} \geq -u_{i1}$. The term corresponding to the first duration time in the likelihood (2.3) must be replaced by the left truncated probability:

$$Pr(Y_{i1}|x_{ij}, Y_{i1} \geq -u_{i1}). \tag{2.9}$$

When the values of u_{i1} are available, then they can be used in (2.9) and adjustments made to likelihood functions like (1.1). However, when they are unknown, then it may be convenient to discard them and treat the process as if the follow-up had begun at time 0, the start of the observation period $(0, M]$. The convenience of this choice depends on whether there are enough within-subject duration times so that this does not represent a substantial loss of information.

Another alternative when $u_{i1} \leq 0$ is to use a model $f_0(u)$ for u_{i1} to provide the marginal distribution for the time Y_{i1} of the first event after selection:

$$f_1(y) = \frac{\int_{-\infty}^{0} f_0(y-u)f_0(u)du}{\int_{-\infty}^{0} S_0(-u)f_0(u)du} \tag{2.10}$$

Care must be taken when chosing a model for u_{i1}. In some cases it is valid to assume $f_0(u) = c$, giving a simplified version of (2.10). A detailed discussion and examples can be found in Cook and Lawless ([15], ch.4) and Lawless and Fong ([33]).

Using conditional models, it is possible to apply the superpopulation and finite population methods described in section 1.5 for each Y_{ij} and the software that has been discussed there can be used for analysis.

Model checking consists of a combination of graphical methods based on residuals and formal tests based on model expansion. The latter involves adding parameters that represent specific types of departures from the current model, and hypothesis tests can be performed. Some examples of model expansion, as discussed in Lawless [32], are: adding covariates representing interactions or nonlinear terms to check a linear model; allowing the scale parameter b in a location-scale model to depend on covariates x as a check on the constancy of b; building time-covariate interactions as a check for the PH assumptions. A detailed discussion on residual and influence analysis as well as model expansion techniques can be found in Lawless [32], and Kalbfleisch and Prentice [25]. The book of Therneau and Grambsch [57], also gives a comprehensive discussion of the model checking methods for the Cox PH model and how they can be implemented in SPlus and SAS.

2.3 Marginal analysis

In some settings researchers might want to study the distribution of single durations in persons that experienced the related event, without covariates. Note that in some studies there might be only a proportion of the population who actually experiences the event of interest in a given period of time. An example is the analysis of the durations of first jobless spells from residents of Ontario, that started in the year 2001. The same applies to durations that can occur more than once in the same individual, for example, all jobless spells that started in 2001.

It is important to distinguish between finite population and superpopulation. When using a finite population model, the statements that are formulated apply to the particular finite population in question. Consider for example, the empirical finite population distribution for all spells that started in Ontario in a particular

year. Suppose that individual i in a finite population $\mathcal{U}$ of size N^* experiences a sequence of jobless spells with durations represented by $\{Y_{i1}, \ldots, Y_{im_i}\}$. The duration distribution as a finite population quantity is expressed as:

$$S_{\mathcal{U}}(y) = \frac{1}{N} \sum_{i \in \mathcal{U}} \sum_{j=1}^{m_i} I(Y_{ij} \geq y). \tag{2.11}$$

where $N = \sum_{i=1}^{N^*} m_i$ is the total number of durations in the population. For individuals i with $m_i = 0$, the corresponding summand in (2.11) is equal to zero.

From the superpopulation framework, we make the assumption that the particular finite population under study represents a realization from a hypothetical superpopulation. That is, the finite population at hand is a member of a set of all possible finite populations in a particular point in time. From this point of view, the empirical distributions for durations from the finite population can approximate a distribution function from the super-population perspective, as the population's size N^* increases to infinity. This is expressed as follows,

$$S(y) = plim\ S_{\mathcal{U}}(y) = plim\ \left(\frac{N^*}{N}\right) plim\ \left(\frac{1}{N^*} \sum_{i \in \mathcal{U}} \sum_{j=1}^{m_i} I(Y_{ij} \geq y)\right).$$

The above is reasonable since the durations Y_{ij} and N are latent random variables at the time the sample is selected, and in this sense, the finite population quantity in (2.11) has random components.

Sometimes we may want to analyze sequences of durations through joint marginal models. The main case is where the potential sequence of durations from each individual has the same length. For example, in the UK Millenium Cohort Study, when studying the times to motor skill developmental events in children, like the time to learn to stand up and the time to learn to jump. These events usually occur around the ages of one and five years, respectively. Suppose that some families dropout from the survey before their children achieve some of these events. The sequence of durations for each individual can be expressed as (Y_{i1}, Y_{i2}) where Y_{i1} denotes the time from birth to learning to stand and Y_{i2} denotes the time from standing to having learned to jump. In studies like this, two issues arise. One is of

induced dependent censoring, when the probability of a second spell to be censored depends on the length of the first. For example, suppose that for some reason a child that has learned to stand up drops out of the study before he learns to jump. The censoring time for Y_{i2} is $C_i - y_{i1}$, that is, is related to the length of time that took him to learn to stand up. The second issue is called non-identifiability, and it arises since we can observe only (Y_{i1}, Y_{i2}) for which $Y_{i1} + Y_{i2} \leq C_{max}$ where $C_{max} = \max\{C_1, \ldots C_n\}$, and C_i is the censoring time for individual i. A detailed discussion can be found in Lin et al., [39] and Cook and Lawless, [15]. This issues make it difficult to apply marginal methods in sequences of durations of variable lengths, and this will not be pursued further in this thesis.

There are several ways to handle within-individual dependence when distributions of duration variables and their relation with covariates are of interest. One way is to introduce subject-specific random effects; this allows for association of duration times within individuals. However, after marginalizing (integrating with respect to random effects distribution), the interpretation of covariate effects can in some cases become awkward. Multivariate models are another alternative (for example, copulas). A third approach is to obtain marginal distributions from conditional models; however, effects of covariates on marginal distributions are generally complex in this setting, except for the normal case. For more details and examples, see Lawless and Fong [33] and Cook and Lawless ([15], in section 4.4.1.).

In the context of marginal modelling from survey data, there are further issues to be considered. An example of a duration study from survey data is when economists are interested in examining permanent layoffs from full-time jobs between 1993-1998. The types of questions that are of interest are: how long does it take a permanently laid-off person to find a new job? What factors determine how long a jobless period lasts? What is the wage gap between a new job and an old one? (Galarneau and Stratychuk [21]). These type of studies have originated the need to develop methodology for marginal regression modelling of durations in the survey context; one example can be found in Kovacevic and Roberts ([29]).

When there is dependency across durations for an individual, a marginal model (one that does not include previous history as covariates) may not be valid for analytical purposes. When an event history process has been in existence before the start of the study, not only event history within the observation period is of relevance, but also information before the start of the study. Again, if dependency is not accounted for, the models may have some descriptive value, but inferences about individual causal factors or dynamics may be misleading. For example, when analyzing first and subsequent jobless spells from 2000 to 2002, it would be useful to have information regarding whether there were any spells before 2000, their length, starting year, etc.

An ideal setup for straightforward interpretation of individual durations is when the individuals in the population are all at risk of experiencing the sequence of spells, and when these sequences have a common starting time across individuals. In individuals who experience two jobless spells in a period of three years, say, it might not make sense to model first spells in the same way as second spells, and it is likely that this kind of modelling would not be of much interest in practice, since the employment experience across individuals most likely has started at different times. Examples of more straightforward interpretation are easier to find in settings where everyone has a common time origin that corresponds to start of follow-up. For instance, in a clinical trial where a certain treatment is administered to all individuals and its effects in a potential sequence of events is monitored over time. In the survey setting, an example would be smoke quitting attempts and their durations in individuals after a set of tobacco preventive measures had been implemented in a given geographical region.

In this thesis, we do not consider joint marginal modelling, instead we focus on univariate marginal modelling and estimation, as well as conditional modelling of sequences of durations.

2.4 Dependent loss to follow-up (LTF)

Dependent loss to follow-up (LTF) is present when there is an association between the LTF mechanism and the event of interest, that is, when LTF depends of event history and also on covariates. For example, when studying unemployment spells, it seems natural to assume that individuals who experience longer spells are more likely to drop out from the survey, than those who have shorter spells. A person may move to a different city as a result of their job search and is not reached by the interviewer, or might feel uncomfortable to stay in the survey while experiencing a long unemployment period and hence refuse to participate.

A feature concerning the study of durations is that these are usually collected intermittently over long periods of time, giving rise to the issue of dependent loss to follow-up. Longitudinal surveys, as discussed previously in section 1.6, usually have this particularity. Loss to follow-up becomes substantial by the end of the observation period, and it may be dependent on events or covariates of interest. In section 1.6 we gave the examples of SLID and the UK MCS study. In the former, samples are typically in the 25-30 percent range of loss to follow-up by the end of the SLID six-year panels. In the MCS study, there was a loss to follow-up rate of 28 percent in the first wave and 42 percent by the second wave (Plewis,[47]). Dependent loss to follow-up has been considered by many authors in the context of continuous and binary outcomes (e.g. Robins et al., [51]; Miller et al., [44]; Preisser et al., [48]); however, duration analysis where data are collected retrospectively has not been considered.

Inverse Probability of Censoring (IPC) weighted methods can be applied to deal with dependent loss to follow-up. These are discussed in the next chapter, and further chapters elaborate on the estimation of marginal duration distributions without covariates and the application of Cox PH regression models to sequences of durations from survey data. This is done taking into account the considerations about conditional and marginal modelling discussed in this chapter. Variance estimation techniques applicable to the K-M estimator and the Cox PH models with

the use of IPC weights are proposed.

Chapter 3

Duration and Event History Analysis with Dependent Loss to Follow-up (LTF)

The main objective of this chapter is to describe some of the proposed methods of analysis in the context of duration and event history analysis, without considering sampling theory for now. These methods have to do with dependent censoring that is caused by an association between the LTF mechanism and the event of interest, that is, when the LTF depends on previous event history and also on covariates. For example, it is more likely for individuals with a higher incidence of unemployment spells to drop out from SLID.

The LTF mechanism manifests itself along the observation period. Information is typically collected at discrete interview times $t = 0, 1, ..., M$. At time t, information for the time interval $(t-1, t]$ is collected. If an individual is lost to follow-up at time $t \leq M$ then we have their data only up to time $t-1$. For a person not lost to follow-up, we have their information over the period $(0, M]$.

The methods to be described involve the use of inverse probability of censoring weights (IPCW) suggested initially by Robins et al. [51]. The first section in this chapter will provide a setup and notation for the IPCW weighting approach and it

is shown how unbiased estimation can be achieved when there is dependent LTF.

Section 3.2 provides a discussion about the modelling method that is employed to estimate the probabilities of dropout. The third section is about the application of the IPCW approach to the analysis of duration time distributions and provides examples regarding employment and unemployment spells like those from SLID, without considering the design features for the moment.

3.1 The Inverse Probability of Censoring Weights (IPCW) method

The IPCW weighting method is designed to give unbiased estimating functions for parameters of interest in the presence of dependent loss to follow-up (Robins et.al [51], Preisser et.al [48]). It consists of modelling the probability of loss to follow-up (LTF) for each individual, at a predetermined set of interview times within the observation period. The IPC weight represents the inverse of the probability of being observed at a given time. This probability is often estimated by a logistic regression model.

There is a second IPC weighting approach discussed in Preisser et al. [48], simpler than the one we use in that the same weight is applied to each individual's duration times, while ours may give one or more weights for a single duration. Preisser et al. used these two approaches for the analysis of longitudinal binary data and illustrated how weighted estimation is consistent when the dropout mechanism is correctly specified. They compare the performance of unweighted and weighted estimating equations under a misspecified dropout model and find that the second weighting approach is less efficient and gives extremely biased estimates under minimal dropout. We also found through simulations that the second approach gives biased results, hence only the first approach will be discussed here.

The probability related to dropout is defined in a set of discrete time points $t = \{0, 1, ..., M\}$ that represent predetermined interview times which are the same

across all individuals. The unemployment spell data from SLID for example, is based on annual interviews for a six year period, in this case $M = 6$. In the discussion below we assume that once a person is lost to follow-up they are not seen again.

The variables to be used are introduced as follows:

$R_t = I(\text{individual is observed at time } t)$ is the indicator related to LTF,

$C = \sup\{t : R_t = 1\}$ is the censoring time, so LTF time is $t + 1$,

$H(t) = \overline{D}(t) = \{D(1), D(2), ..., D(t), H(0)\}$ is the history of the process up to time t, where H(0) represents the initial conditions measured at $t = 0$ and $D(t)$ is data over the interval $(t - 1, t]$ including covariates, collected at time t.

Let us assume that the event of being observed at time t is unrelated to the current and future outcomes and covariates, conditional on the observed past. This is the missing at random (MAR) assumption, in the sense of Rubin [53]. That is:

$$Pr(R_t = 1|R_{t-1} = 1, H(M)) = Pr(R_t = 1|R_{t-1} = 1, H(t-1)). \tag{3.1}$$

Consider the model $P(D(t)|Z(t))$ where $Z(t)$ can include external covariates $X(t)$ plus history $H(t - 1)$ up to the previous observation time. The likelihood is expressed by pieces according to $H(M) = \{D(1), ..., D(M), H(0)\}$, given external covariate history $X(M)$.

Then the score function for the ith individual is given by:

$$U_i^R(\theta) = \sum_{t=1}^{M} R_{it} \partial \log P_\theta[D_i(t)|Z_i(t)]/\partial\theta = \sum_{t=1}^{M} u_{it}^R(\theta).$$

If R_t and $D(t)$ are independent given $Z(t)$ $(R_t \perp D(t)|Z(t))$ then the above can be used directly to obtain unbiased estimates of θ provided the model $P_\theta(D_i(t)|Z_i(t))$ is correctly specified so that $E\left\{u_{it}^R\right\} = 0$ for each $t = 1, 2, ..., M$. It is verified that:

$$\begin{aligned} E_{D_i(t),R_{it}|Z_i(t)}\left\{u_{it}^R\right\} &= E_{D_i(t)|Z_i(t)}\left\{E_{R_{it}|Z_i(t)}\left\{u_{it}^R(\theta)\right\}\right\} \\ &= E_{D_i(t)|Z_i(t)}\left\{\partial \log P_\theta[D_i(t)|Z_i(t)]/\partial\theta\right\} E_{R_{it}|Z_{it}}\left\{R_{it}\right\} = 0. \end{aligned} \tag{3.2}$$

However, if $R_t \perp D(t)|Z(t)$ is not true but variables $Z^c(t)$ which include $Z(t)$ are available such that $R_t \perp D(t)|Z^c(t)$, then the following estimating function is appropriate:

$$U_i^p(\theta) = \sum_{t=1}^{M} \frac{R_{it}}{p_{it}} \frac{\partial \log P_\theta[D_i(t)|Z_i(t)]}{\partial \theta} = \sum_{t=1}^{M} u_{it}^p(\theta), \tag{3.3}$$

where $p_{it} = Pr(R_{it} = 1|Z_i^c(t)) = Pr(C_i \geq t|Z_i^c(t))$. Under the correct model, unbiasedness can be verified by:

$$\begin{aligned}
E_{D_i(t),R_{it},Z_i^c(t)|Z_i(t)}\left\{u_{it}^p(\theta)\right\} &= E_{D_i(t),Z_i^c(t)|Z_i(t)}\left\{E_{R_{it}|D_i(t),Z_i^c(t),Z_i(t)}\left\{u_{it}^p(\theta)\right\}\right\} \\
&= E_{D_i(t),Z_i^c(t)|Z_i(t)}\left\{E_{R_{it}|Z_i^c(t)}\left\{u_{it}^p(\theta)\right\}\right\} \qquad (3.4) \\
&= E_{D_i(t)|Z_i(t)}\left\{\partial \log(P_\theta[D_i(t)|Z_i(t)])/\partial\theta\right\} \qquad (3.5) \\
&= 0. \qquad (3.6)
\end{aligned}$$

Note that (3.4) is due to the fact that $Pr(R_t|D(t), Z^c(t), Z(t)) = Pr(R_t|Z^c(t))$ since $Z_i^c(t)$ includes $Z_i(t)$ and the assumption that $R_t \perp D(t)|Z^c(t)$.

Line (3.5) follows by letting $G(D(t)|Z(t)) = \partial \log(P_\theta[D_i(t)|Z_i(t)])/\partial\theta$ and

$$\begin{aligned}
E_{D(t),Z^c(t)|Z(t)}\left\{G(D(t)|Z(t))\right\} &= \int_{D(t)} \int_{Z^c(t)} G(D(t)|Z(t)) dP(D(t), Z^c(t)|Z(t)) \\
&= \int_{D(t)} G(D(t)|Z(t)) dP(D(t)|Z(t)) \\
&= E_{D(t)|Z(t)} G(D(t)|Z(t)). \qquad (3.7)
\end{aligned}$$

It can be shown that if

$$Pr\big(D_i(t)|Z_i^c(t)\big) = Pr\big(D_i(t)|Z_i(t)\big) \text{ then } Z_i^c(t) \perp D_i(t)|Z_i(t),$$

and so $D_i(t) \perp R_i(t)|Z_i(t)$. Thus the IPC weighting is needed only when there are covariates or event history that affect both $R_i(t)$ and $D_i(t)$, but which are not in the model $Pr\big(D_i(t)|Z_i(t)\big)$. This can often occur, in particular when we are interested in marginal distributions for durations or other responses, or distributions that condition on just a few covariates.

By noting that $P_\theta(H_i(C_i)) = \prod_{t=1}^{C_i} P_\theta(D(t)|Z(t))$, the estimating function in (3.3) is equivalent to:

$$U_i^p(\theta) = \sum_{t=1}^{C_i} \frac{1}{p_{it}} \frac{\partial log P_\theta(D_i(t)|Z_i(t))}{\partial \theta}.$$

This weighting approach needs a separate weight for the data from each year, so the data sets must be arranged accordingly for analysis. This will be discussed in section 3.3.

3.2 Modelling the dropout process

This section elaborates more explicitly on the dropout modelling that is going to be used to provide the weights in expression (3.3). The estimation of dropout probabilities from the modelling approaches defined in the preceding section are carried out using the following logistic model:

$$\text{logit } \lambda_{it}(Z_i^c(t); \alpha) = \alpha' Z_i^c(t), \tag{3.8}$$

where $\lambda_{it}(Z_i^c(t); \alpha) = Pr(R_{it} = 1|R_{i,t-1} = 1, Z_i^c(t); \alpha)$ is the probability that individual i was observed at interview time t given that they were observed at $t-1$, $t = 1, ..., M$. Also, $Z_i^c(t)$ is a $p \times 1$ covariate vector and α is a $p \times 1$ parameter vector. As mentioned earlier, for the IPCW to be needed, the covariates $Z_i^c(t)$ must affect both the durations and the dropout process.

Under the assumption that $Pr(R_{it} = 1|R_{i,t-1} = 1, Z_i^c(t)) = Pr(R_{it} = 1|R_{i,t-1} = 1, H_i(t-1)) = Pr(R_{it} = 1|R_{i,t-1} = 1, H_i(M))$, estimates for the LTF probabilities can be obtained by using the fitted coefficients from (3.8) and by

$$Pr(R_{it} = 1|Z_i^c(t)) = \lambda_{i1} \cdot \lambda_{i2} \cdots \lambda_{it}. \tag{3.9}$$

It is assumed that $R_{i1} = 1$ with probability 1 and that intermittent dropout patterns are not allowed: $R_{i,t+k} = 0, k > 0$ whenever $R_{it} = 0$. The likelihood

function for the dropout probability in (3.8) is given by:

$$L(\alpha) = \prod_{t=1}^{M} \prod_{i=1}^{m} \left\{ \lambda_{it}(Z_i^c(t);\alpha)^{R_{it}} (1 - \lambda_{it}(Z_i^c(t);\alpha))^{1-R_{it}} \right\}^{R_{i,t-1}}. \tag{3.10}$$

The dropout probabilities in (3.10) can be estimated for the discrete interview times $t = 1, ..., M$, via standard GLM software (Lawless (2003), p.372).

3.3 Examples

This section provides an example of the discussion about the application of the IPCW approach from section 3.1 to the analysis of duration time distributions and event histories.

Suppose we have an event history process over a time period $(0, M]$ represented by a model with parameter of interest θ and possibly other nuisance parameters. Let this event history process consist of sojourns and transitions within a set of states. Since censoring may be present in this process, we will consider notation relative to an observed sequence of sojourns over $(0, C]$, where $C \leq M$.

For illustration, suppose that for all individuals the process starts in the same state and the subsequent visited states are the same for everyone. Let j indicate the sojourn in state $E_j \in \{1, 2, ..., K\}$ which starts at time t_{j-1} and ends at time t_j, $j = 1, 2, ...$ (let $t_0 = 0$). Suppose $\delta_j = 1$ indicates if the observed sojourn ends at t_j with a transition to a new state and $\delta_j = 0$ if it ends due to censoring with the individual remaining in state E_j. Finally, let Y_j be the full duration of the jth sojourn, so that $y_j = t_j - t_{j-1} \leq Y_j$ is the observed sojourn.

Note that the process occurs in continuous time while the information is collected at discrete times $t \in \{1, 2, ..., M\}$. Also, let x denote the history of the external covariates, which can be time varying. In this case it is assumed that the transition intensities at time t will depend only on covariate values up to time t. To keep things simple we assume that given $H(0)$, the sequence of states that can be visited is known.

The probability models to represent the whole sequence of sojourns may be represented by:

$$\prod_j Pr(Y_j|Y^{(j-1)}, x, H(0)) = Pr(Y_1, Y_2, ...|x, H(0)), \tag{3.11}$$

where $Y^{(j-1)} = \{Y_1, ..., Y_{j-1}\}$. Note that (3.11) allows us to consider either conditional or joint model specifications for the durations $Y_1, Y_2, ...$

If censoring is present, then the conditional specification in (3.11) is more convenient to use. Assume that censoring is ignorable and that the process for individual i was observed over the time interval $(0, C_i]$. The likelihood function for a sequence of m_i observed durations $y_{i1}, ..., y_{i,m_i}$ (the last one right censored) based on the conditional model in (3.11) is

$$L_i(\theta) = \prod_{j=1}^{m_i-1} f_{ij}(y_{ij}|y_i^{(j-1)}x_i, H_i(0)) \cdot S(y_{i,m_i}|y_i^{(m_i-1)}, x_i, H_i(0)), \tag{3.12}$$

where $y_i^{(m_i-1)} = \{y_{i1}, ..., y_{i,m_i}\}$, $f(\cdot)$ and $S(\cdot)$ are density and survivor functions.

The estimating function based on (3.12) for an individual i is $u_i(\theta) = \partial \log L_i(\theta)/\partial\theta$ and across n independent individuals is

$$U(\theta) = \sum_{i=1}^{n} U_i(\theta). \tag{3.13}$$

If censoring is not ignorable given the external covariates x_i, then the IPCW approach discussed in section 3.1 can be used. Let $U_{it}^p(\theta)$ represent the term $\partial \log P_i(D_i(t)|Z_i(t))/\partial\theta$ in expression (3.3) for $t = 1, 2, ..., M$. The estimating function is:

$$\begin{aligned} U^p(\theta) &= \sum_{i=1}^{n}\sum_{t=1}^{M} \frac{R_{it}}{p_{it}} U_{it}^p(\theta) \\ &= \sum_{i=1}^{n} U_i^p(\theta). \end{aligned} \tag{3.14}$$

The data $D_i(t)$ over the time interval $(t-1, t]$ correspond to the durations Y_{ij} observed in $(t-1, t]$, while the set of covariates $Z_i(t)$ contains relevant information

in $H_i(t-1)$ about the event process and covariates. If the duration Y_{ij} started before time $t-1$, then $H_i(t-1)$ contains information about the starting time of the related spell.

For example, suppose that at time t, a sojourn duration Y_{ij} has length $y_{ij}(t)$. Then, the information contained in $H_i(t-1)$ will be $y_{ij}(t-1)$ and the information $D(t)$ includes either that the spell ends in $(t-1,t]$ (due to a transition or censoring) or extends beyond t. So the likelihood contribution from $D(t)$ for an individual i with such a spell is of the form:

$$Pr(Y_{ij} = y_{ij}|Y_{ij} > y_{ij}(t-1), z_{ij})^{\delta_{ij}(t)} \cdot Pr(Y_{ij} > y_{ij}(t-1)+1|Y_{ij} > y_{ij}(t-1), z_{ij})^{1-\delta_{ij}(t)},$$

where $\delta_{ij}(t) = I(Y_{ij}$ ends in $(t-1,t])$.

Since each data collection interval $(t-1,t]$ is dealt with separately for $t = 1,2,...,M$, the estimating function obtained from the model must be unbiased over each interval, that is, the basic estimating functions $U_{it}^{p}(\theta)$ in (3.14) need to be unbiased for each t . This should be considered carefully if the model accounts only for partial information about the previous event history.

The UK Millennium Study has interview times of 9 months, 3, 5 and 7 years. An example of the kind of sequences of durations described above might be related to the times to cognitive developmental milestones in children up to seven years of age. Suppose that $K = 4$ and $E = 1$ if the child turns his head when hearing his name (6-9 months), $E = 2$ when the child can match two objects together by color, shape or size (1-2 years), $E = 3$ when he learns different shapes by name and colors (3-5 years) and $E = 4$ when the child learns his full name, age and address (5-7 years). In this example, it may be reasonable to assume that the time to achieve one of these milestones is related to the time that previous ones took to occur. Further, censoring is likely to be present as well as dependent loss to follow-up as there might exist factors that affect both the development of a child and his loss to follow-up, such as socio-economic status and family composition.

Alternation between employment and unemployment

As another example, we consider a sequence of four transitions between states $E_j \in \{E, U\}$ where E and U stand for employed and unemployed, respectively. Recall that Y_{ij} is the full duration of the jth sojourn, and the observed duration is $y_{ij} = t_{ij} - t_{i,j-1} \leq Y_{ij}$ where $t_{i,j-1}$ and t_{ij} denote the starting and ending times of the observed sojourn, respectively. Suppose that the individual's sequence started with the state E. Let C_i denote the time that the individual i was last seen, $C_i \in \{1, 2, ..., M\}$.

Suppose that the hazard functions used to describe the unemployment and employment durations are given, respectively, by:

$$\lambda_U(y_j|x, y^{(j-1)}, H(0); \theta) \text{ and } \lambda_E(y_j|x, y^{(j-1)}, H(0); \psi).$$

Further, assume that the times of $E \leftrightarrow U$ transitions $\{t_{i1}, t_{i2}, t_{i3}, t_{i4}\}$ and sojourn durations $\{y_{i1}, y_{i2}, y_{i3}, y_{i4}\}$ with y_{i4} censored, were recorded. Note that the corresponding states are (see figure 3.1):

$$\{E_{i1} = E,\ E_{i2} = U,\ E_{i3} = E,\ E_{i4} = U\}.$$

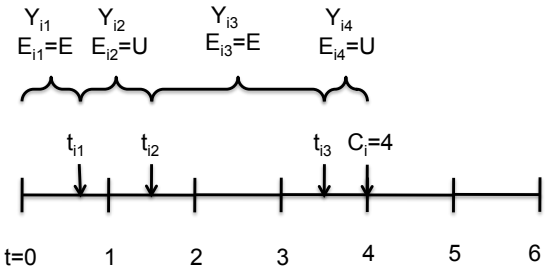

Figure 3.1: Example of a sequence of employment and unemployment durations.

Assuming that there is no information on the starting time of the first E duration

Y_{i1}, the likelihood for years $t = 1, 2, 3, 4$ is composed of the following probabilities:

$$\begin{aligned}
&\text{For } D(1): \quad Pr(Y_{i1} = y_{i1}|z_{i1})Pr(Y_{i2} > 1 - y_{i1}|y_{i1}, z_{i1}),\\
&\text{For } D(2): \quad Pr(Y_{i2} = y_{i2}|Y_{i2} > 1 - y_{i1}, z_{i2})Pr(Y_{i3} > 2 - t_{i2}|y_{i1}, y_{i2}, z_{i2}),\\
&\text{For } D(3): \quad Pr(Y_{i3} > 3 - t_{i2}|y_{i1}, y_{i2}, z_{i3}),\\
&\text{For } D(4): \quad Pr(Y_{i3} = y_{i3}|Y_{i3} > 3 - t_{i2}, z_{i4})Pr(Y_{i4} > 4 - t_{i3}|y_{i1}, y_{i2}, y_{i3}, z_{i4}).
\end{aligned}$$

A data frame to implement this approach is illustrated in Table 3.1. As mentioned above, the idea is to partition the time interval $[0, 6]$ into pieces defined by the transitions to U/E states and the interview times $t = 1, 2, ..., 6$. Note that a single individual with **ID**=i will have as many lines as pieces that resulted in his or her event history in their follow-up period. The individual in our example has 7 pieces altogether, as Figure 3.1 shows. Therneau and Grambsch [57] refer to this as "counting process data" format.

The information provided in each line in the data frame will correspond to the event process and covariate information available for each interval. The variables **Start.t** and **Stop.t** give the calendar times that define the starting and ending times of the intervals. The variable **Start.y** is zero to indicate that a new spell begun at Start.t, otherwise it will show the length of the spell at the end of the previous interval. For example, the third line corresponds to the interval given by (Start.t, Stop.t)=$(1, t_{i2})$ and has a Start.y value of $1 - t_{i1}$, which is the length of the spell at the end of the previous interval: (Start.t, Stop.t)=$(t_{i1}, 1)$. **Stop.y** gives the cumulative length of the spell at the end of the corresponding interval from the time it started. For example, lines 4,5 and 6 give the length of the spell Y_{i3} at the end of Stop.t=$2, 3, t_{i3}$. **Etype** gives the type of transition (U or E) that occurred in the interval to which each line corresponds. **Status** is 1 if the spell ended by Stop.t, 0 if it extended beyond Stop.t. The elements of the **Covs** column give information on previous lengths as well as external covariates. Note that the information in this column coincides with the covariates and previous lengths used for conditioning in the duration probabilities for $D(1), ..., D(4)$ described above. Finally, note that the

ID	Weight	Start.t	Stop.t	Start.y	Stop.y	Etype	Status	Covs*	Enum
i	p_{i1}^{-1}	0	t_{i1}	0	t_{i1}	E	1	y_{i1}, z_{i1}	1
i	p_{i1}^{-1}	t_{i1}	1	0	$1-t_{i1}$	U	0	$y_i^{(3)}, z_{i2}$	2
i	p_{i2}^{-1}	1	t_{i2}	$1-t_{i1}$	$t_{i2}-t_{i1}$	U	1	$y_i^{(4)}, z_{i3}$	3
i	p_{i2}^{-1}	t_{i2}	2	0	$2-t_{i2}$	E	0	$y_i^{(4)}, z_{i4}$	4
i	p_{i3}^{-1}	2	3	$2-t_{i2}$	$3-t_{i2}$	E	0	$y_i^{(4)}, z_{i4}$	5
i	p_{i4}^{-1}	3	t_{i3}	$3-t_{i2}$	$t_{i3}-t_{i2}$	E	1	$y_i^{(4)}, z_{i4}$	6
i	p_{i4}^{-1}	t_{i3}	4	0	$4-t_{i3}$	U	0	$y_i^{(4)}, z_{i4}$	7

* Note: $y_i^{(k)} = (y_{i1}, ..., y_{i,k-1})$

Table 3.1: Data frame to implement IPCW methods.

Weight column gives values of p_{it}^{-1} to those intervals (lines) that are related to the time intervals $(t-1, t]$, $t = 1, 2, 3, 4$.

Chapter 4

Weighted Parametric Regression Analysis

This chapter provides estimation methods to implement the IPCW techniques from Chapter 3 in the context of duration variables from survey data. The techniques are based on estimating function theory (White [58]) on parametric regression models. This chapter also constitutes the basis for the methods that are proposed in chapters 5 and 6, where Kaplan-Meier and Cox PH models are discussed.

Robins et al. [51] gave variance estimation methods that can be derived from those of White, applied to IPCW generalized linear models for longitudinal data in the non-survey context. Their "sandwich" variance estimators take into account the random nature of the weights. Miller et al. [44] have extended the methods from Robins et al. to survey data. They analyze discrete outcomes from longitudinal surveys subject to multiple-cause non-response, accommodate for sampling design weights and use a stratified cluster-sampling version of the middle part of the "sandwich" variance estimate from Robins et al. The structure of duration and event history data requires modifications and extensions of the previous methods, and we develop these here and in the following chapters.

Section 4.1 provides a brief overview of variance estimation in estimating function theory and gives an adaptation of the IPCW method from Robins et al. to

the parametric duration analysis framework. An example is provided regarding location-scale models together with a simulation in which the behavior of the IPCW based methodology is examined. Robins et al. show that the asymptotic variance of the estimator of the parameters of interest are smaller when estimated IPC weights are used instead of pre-specified weights. In addition, naively treating the IPC weights as fixed gives slightly larger variance estimates, but they are often close to the variances obtained when treating them as random. The simulations from this chapter are used to compare the variance estimates based on random or non-random assumptions for the IPC weights. Section 4.2 gives an extension of the variance estimates from Robins et al. to the analysis of durations from survey data, along the lines of Miller et al.

4.1 Estimation in classical settings

The following discussion regards estimation using IPC weights in the parametric analysis of right-censored data, along the lines of Robins et al. [51]. Let $(0, M]$ denote the follow-up period for individuals $i = 1, 2, ..., n$ and suppose $m_i \geq 0$ events are observed for individual i. Let u_{ij} and v_{ij} denote the start and end times of a spell with an associated duration defined by $Y_{ij} = v_{ij} - u_{ij},\ j = 1, 2, ..., m_i$. Let the time the person was last seen be $C_i \in \{1, 2, ..., M\}$. Note that this implies that an individual may be lost to follow-up before the end of the period $(0, M]$.

Furthermore, suppose that $Z_{ij}(u_{ij} + y)$ denotes a set of covariates for individual i, at calendar time $u_{ij} + y$; it may include u_{ij} and also may include information on prior event history up to time $t - 1$, where $t - 1 < u_{ij} + y \leq t$.

Let $S(y)$, $h(y)$ and $f(y)$ denote the survivor, hazard, and probability density functions for a specific duration time y given covariates where $S(\cdot)$, $h(\cdot)$ and $f(\cdot)$ depend on a finite dimensional parameter θ.

The discussion from Chapter 3 presents two models for estimation, one for modelling the events or durations of interest and another one for modelling loss

to follow-up to obtain the IPC weights. They lead to the following system of estimating equations:

$$\begin{aligned} U(\theta, \alpha) &= 0 \\ G(\alpha) &= 0 \end{aligned} \tag{4.1}$$

where θ and α are parameter vectors of dimension px1 and qx1. The estimating functions to model durations have the following form:

$$U(\theta, \alpha) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{t=1}^{M} w_{it}(y) U_{ijt}(\theta) = \sum_{i=1}^{n} U_i(\theta), \tag{4.2}$$

where $w_{it}(\alpha) = R_{it}/p_{it}(\alpha)$ and $U_{ijt}(\theta) = \partial l_{ijt}(\theta)/\partial\theta$. In our case, the term $l_{ijt}(\theta)$ has the form of the log-likelihood of a survival model for right censored data. IPC weights are associated to time intervals $(t-1, t]$ and durations can extend over one or more of these. Therefore, the duration model in (4.1) needs to take into account the delayed entry of spells that start before a given interval $(t-1, t]$. Let

$$y_{ij}(t) = \min(t, v_{ij}) - \min(t, u_{ij})$$

be the length of the observed duration y_{ij} at time t and

$$\delta_{ij}(t) = I(t - 1 < v_{ij} \leq t)$$

indicate whether duration y_{ij} ends in the interval $(t-1, t]$; the log-likelihood $l_{ijt}(\theta) = \log L_{ijt}(\theta)$ has the form:

$$L_{ijt}(\theta) = \left\{ \frac{f(y_{ij})}{S(y_{ij}(t-1))} \right\}^{\delta_{ij}(t)} \left\{ \frac{S(y_{ij}(t))}{S(y_{ij}(t-1))} \right\}^{1-\delta_{ij}(t)}. \tag{4.3}$$

The estimating function related to loss to follow-up is the score function for a logistic model for the probability of being observed at time t, given that $R_{i,t-1} = 1$, and is given by:

$$G(\alpha) = \sum_{i=1}^{n} \sum_{t=1}^{M} (R_{it} - \lambda_{it}(\alpha) R_{i,t-1}) \frac{\partial \operatorname{logit} \lambda_{it}(\alpha)}{\partial \alpha} = \sum_{i=1}^{n} G_i(\alpha). \tag{4.4}$$

where

$$\begin{aligned} \operatorname{logit}\ (\lambda_{it}(\alpha)) &= \operatorname{logit} \{\Pr(R_{it} = 1 | R_{i,t-1} = 1, Z_i^c(t); \alpha\} \\ &= \alpha_t' Z_i^c(t), \end{aligned} \tag{4.5}$$

and α_t is a vector of regression coefficients, $\alpha = (\alpha_1', \ldots, \alpha_M')'$, and $Z_i^c(t)$ a set of covariates that may affect both durations and dropout. Note that $p_{it}(\alpha) = \Pr(R_{it} = 1 | Z_i^c(t)) = \lambda_{i1}(\alpha) \ldots \lambda_{it}(\alpha)$. Let the dimension of α_t be q_t and so the vector $\alpha = (\alpha_1', \ldots, \alpha_M')'$ has dimension $q = q_1 + \cdots + q_M$. The model (4.4) can be fitted with standard logistic regression or generalized linear model software to give maximum likelihood estimates $\hat{\alpha}_t$ and estimated probabilities $\hat{p}_{it} = p_{it}(\hat{\alpha})$.

Note that the components of the estimating function for loss to follow-up $G(\alpha)$ in (4.4) come from the separate logistic regression model loglikelihood functions for $t \in \{1, \ldots, M\}$. For a given value of t this is

$$\sum_{i=1}^{n} (R_{it} - \lambda_{it}(\alpha) R_{i,t-1}) \frac{\partial \operatorname{logit} \lambda_{it}(\alpha)}{\partial \alpha}. \tag{4.6}$$

The variance estimate for $\hat{\theta}$ comes from a direct application of the results of White [58] on estimating function theory. The estimate of the asymptotic covariance matrix for the parameter $\hat{\psi} = (\hat{\theta}', \hat{\alpha}')'$ is consistent and robust for model misspecification, and is given by

$$\widehat{Var}(\hat{\psi}) = A(\hat{\psi})^{-1} B(\hat{\psi})^{-1} A(\hat{\psi})^{-1}, \tag{4.7}$$

where

$$A(\psi) = \begin{pmatrix} -\partial U(\theta, \alpha) / \partial \theta' & -\partial U(\theta, \alpha) / \partial \alpha' \\ -G(\alpha) / \partial \theta' & -\partial G(\alpha) / \partial \alpha' \end{pmatrix} = \begin{pmatrix} A_{11}(\theta, \alpha) & A_{12}(\theta, \alpha) \\ 0 & A_{22}(\alpha) \end{pmatrix}$$

$$B(\psi) = \begin{pmatrix} Var(U(\theta, \alpha)) & Cov(U(\theta, \alpha), G(\alpha)) \\ Cov(G(\alpha), U(\theta, \alpha)) & Var(G(\alpha)) \end{pmatrix} = \begin{pmatrix} B_{11}(\theta, \alpha) & B_{12}(\theta, \alpha) \\ B_{21}(\theta, \alpha) & B_{22}(\alpha) \end{pmatrix}$$

The estimate of the variance of $\hat{\theta}$ is then obtained as the upper left block of (4.7) evaluated at $(\hat{\theta}, \hat{\alpha})$,

$$\begin{aligned} \widehat{Var}(\hat{\theta}) &= A_{11}(\hat{\theta}, \hat{\alpha})^{-1} \left\{ B_{11}(\hat{\theta}, \hat{\alpha}) - A_{12}(\hat{\theta}, \hat{\alpha}) A_{22}(\hat{\alpha})^{-1} B_{21}(\hat{\theta}, \hat{\alpha}) \right\} A_{11}(\hat{\theta}, \hat{\alpha})^{-1}, \\ &\cong^a A_{11}(\hat{\theta}, \hat{\alpha})^{-1} \left\{ B_{11}(\hat{\theta}, \hat{\alpha}) - B_{12}(\hat{\theta}, \hat{\alpha}) B_{22}(\hat{\alpha})^{-1} B_{21}(\hat{\theta}, \hat{\alpha}) \right\} A_{11}(\hat{\theta}, \hat{\alpha})^{-1}. \end{aligned} \tag{4.8}$$

The asymptotic equivalence of

$$E\{A_{22}(\alpha)\} \text{ and } \{B_{22}(\alpha)\},$$

and of

$$E\{A_{12}(\theta,\alpha)\} \text{ and } E\{B_{12}(\theta,\alpha)\}$$

allows to replace the corresponding terms in $\widehat{Var}(\hat{\theta})$, expression (4.8). This asymptotic equivalence can be shown noting that,

1. $E\{A_{22}(\alpha)\} = E\{-\partial G(\alpha)/\partial\alpha'\} = Var\ (G(\alpha)) = \{B_{22}(\alpha)\}$ since $G(\alpha)$ is based on likelihood functions.

2. $E\{A_{12}(\theta,\alpha)\} = E\{B_{12}(\theta,\alpha)\}$ since R_{it} is assumed to be conditionally independent of the entire duration history $H(M)$, given covariates $Z_i^c(t)$. The i'th terms of $A_{12}(\theta,\alpha)$ are $(A_{12i})_s = -\partial U_i(\theta,\alpha)/\partial\alpha_s$,

$$\begin{aligned}(A_{12i})_s &= I(m_i>0)\sum_{t=1}^{M}\sum_{j=1}^{m_i} w_{ijt}(y)\frac{\partial \log p_{it}(\alpha)}{\partial\alpha_s}\frac{\partial l_{ijt}(\theta)}{\partial\theta} \\ &= I(m_i>0)\sum_{j=1}^{m_i}\sum_{t=1}^{M} I(s\le t)w_{ijt}(y)Z_i^c(s)\left[1-\lambda_{is}(\alpha_s)\right]\frac{\partial l_{ijt}(\theta)}{\partial\theta}\end{aligned}$$

for $s = 1,\ldots,M$, since under the logistic model in (4.4),

$$\begin{aligned}\frac{\partial \log\left(p_{it}(\alpha)\right)}{\partial\alpha_s} = \sum_{s'\le t}\frac{\partial \log\lambda_{is'}(\alpha_{s'})}{\partial\alpha_s} &= I(s\le t)\frac{\partial\log\lambda_{is}(\alpha_s)}{\partial\alpha_s} \\ &= I(s\le t)Z_i^c(s)\left[1-\lambda_{is}(\alpha_s)\right]\end{aligned}$$

In addition, $E\{(B_{12i})_s\} = E\{U_i(\theta,\alpha)\ G_i(\alpha)_s\} = 0$ for $s > t$, since R_{is} is independent of the entire duration history $H_i(M)$, conditional on $Z_i^c(s)$.

For $s \le t$, we have $E\{R_{is} - \lambda_{is}(\alpha_s)R_{i,s-1}|R_{i,s-1}=1, R_{it}=1\} = 1-\lambda_{is}(\alpha_s)$. Thus,

$$\begin{aligned}E\left\{(B_{12i})_s\right\} &= E\left\{I(m_i>0)\sum_{j=1}^{m_i}\sum_{t=1}^{M} w_{ijt}(y)Z_i^c(s)\left\{R_{is}-\lambda_{is}(\alpha_s)R_{i,s-1}\right\}\frac{\partial l_{ijt}(\theta)}{\partial\theta}\right\} \\ &= E\left\{I(m_i>0)\sum_{j=1}^{m_i}\sum_{t=1}^{M} I(s\le t)w_{ijt}(y)Z_i^c(s)\left[1-\lambda_{is}(\alpha_s)\right]\frac{\partial l_{ijt}(\theta)}{\partial\theta}\right\} \\ &= E\left\{(A_{12i})_s\right\}.\end{aligned}$$

The second line in expression (4.8) is equivalent to the variance estimate for $\hat{\theta}$, expressed by Robins et al. [51] as:

$$\widehat{Var}(\hat{\theta}) = \hat{B}^{-1}\hat{C}\hat{B}^{-1\,\prime}, \tag{4.9}$$

where both $\hat{B}$ and $\hat{C}$ are *pxp* matrices,

$$\hat{B} = -\frac{1}{\sqrt{n}}\frac{\partial U(\theta,\hat{\alpha})}{\partial \theta}\Big|_{\hat{\theta}} \text{ and} \tag{4.10}$$

$$\hat{C} = \frac{1}{n}\sum_{i=1}^{n}\left\{U_i(\hat{\theta},\hat{\alpha}) - \left[\sum_{i=1}^{n}U_i(\hat{\theta},\hat{\alpha})G_i(\hat{\alpha})'\right)\right]\left[\sum_{i=1}^{n}G_i(\hat{\alpha})G_i(\hat{\alpha})'\right]^{-1}G_i(\hat{\alpha})\right\}^{\otimes 2} \tag{4.11}$$

where $A^{\otimes 2} = AA'$. It can be readily shown that the expression for $\hat{C}$ in (4.11) disregarding the term $1/n$ is equivalent to the middle term in the second line of (4.8). That is,

$$\sum_{i=1}^{n}\left\{U_i(\hat{\theta},\hat{\alpha}) - \left[\sum_{i=1}^{n}U_i(\hat{\theta},\hat{\alpha})G_i(\hat{\alpha})'\right)\right]\left[\sum_{i=1}^{n}G_i(\hat{\alpha})G_i(\hat{\alpha})'\right]^{-1}G_i(\hat{\alpha})\right\}^{\otimes 2}$$
$$= B_{11}(\hat{\theta},\hat{\alpha}) - B_{12}(\hat{\theta},\hat{\alpha})B_{22}(\hat{\theta},\hat{\alpha})^{-1}B_{21}(\hat{\theta},\hat{\alpha})$$

where

$$B_{11}(\hat{\theta},\hat{\alpha}) = \sum_{i=1}^{n}U_i(\hat{\theta},\hat{\alpha})U_i(\hat{\theta},\hat{\alpha})' \tag{4.12}$$

$$B_{12}(\hat{\theta},\hat{\alpha}) = \sum_{i=1}^{n}U_i(\hat{\theta},\hat{\alpha})G_i(\hat{\alpha})' \tag{4.13}$$

$$B_{22}(\hat{\theta},\hat{\alpha}) = \sum_{i=1}^{n}G_i(\hat{\alpha})G_i(\hat{\alpha})' \tag{4.14}$$

Note that the matrix $B_{22}(\hat{\alpha})$ in (4.8) can be replaced by a block diagonal matrix representing $Var(G(\alpha))$, since the estimating functions in (4.4) are estimated separately for each value of t and are mutually independent (see expression (4.6)). The use of a block diagonal version simplifies the computation of its inverse. Moreover, note that the summands of the $\hat{C}$ matrix have the form of square cross products of the residuals from the multivariate regression of the $U_i(\theta,\hat{\alpha})$ vectors on the vectors

$G_i(\hat{\alpha})$ and thus they can be computed using standard linear models software. For instance, using the `lm` function in R/SPlus and then using the `$residuals` option .

Robins et al. point out that augmenting a correctly specified model for loss to follow-up usually leads to an improvement in the efficiency with which θ is estimated and show that there exists a lower bound for the asymptotic variance of $\hat{\theta}$. One of the arguments for this gain in efficiency is that the variance matrix of the residuals from a multivariate regression decreases as the number of covariates increases. Furthermore, Robins et al. state that the variance of $\hat{\theta}$ using (4.9) is larger than the variance computed considering the weights $p_{it}^{-1}(\hat{\alpha})$ as fixed. This is seen for the alternative variance estimate (4.8), by noting that (4.8) with the second term in the middle is dropped, is the variance estimate for $\hat{\theta}$ when the weights are known.

An example: location-scale models

For illustration, consider the family of location-scale or accelerated failure time regression models and suppose that we have a single duration for each subject, denoted by y_i. Assume time varying covariates that may include information related to the durations are constant over intervals $(t-1, t]$, denoted by $Z_i(t)$. Consider $y_i^* = \log(y_i)$ and location and scale parameters $\mu_{it} = Z_i(t)'\beta$ and b, respectively, where β is the parameter vector of regression coefficients.

Defining $\theta = (\beta, b)$ and $\bar{y}_i = (\log(y_i) - Z_{it}'\beta)/b$, (4.3) gives (for $m_i = 1$):

$$l_{it}(\theta) = \log\left\{\left[\frac{f_0(\bar{y}_i)/b}{S_0(\bar{y}_i(t-1))}\right]^{\delta_{it}}\left[\frac{S_0(\bar{y}_i(t)))}{S_0(\bar{y}_i(t-1))}\right]^{1-\delta_{it}}\right\} \tag{4.15}$$

where $y_i(t) = \min(t, v_i) - \min(t, u_i)$ is the length of the spell y_i up to time t, $t = 0, 1, ..., M$. Also, $f_0(\cdot)$, $h_0(\cdot)$ and $S_0(\cdot)$ are the respective density, hazard and survivor functions of $\bar{y}_i$. Some models that could be used for $f_0(y)$ are the extreme-value, normal and logistic, which correspond to y_i being Weibull, log-normal, and

log-logistic.

As before, the IPC weights are obtained by solving the estimating function for loss to follow-up, $G(\alpha)$ in (4.4), for separate logistic regression models as in (4.6).

In order to compute the variance estimates we need the matrix of individual contributions to the duration model scores $U_i(\theta)$ and also the matrix of individual contributions to the loss to follow-up model scores $G_i(\alpha)$. In parametric location-scale modelling this can be done in several ways.

The estimate for β in equation (4.2) can be obtained by using general optimization software, that is, by maximizing the weighted log-likelihood function which corresponds to

$$l_w(\theta, \alpha) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{t=1}^{M} w_{it}(y) l_{ijt}(\theta)$$

where $l_{ijt}(\theta)$ is of the form (4.15). For example, (i) in R via the `nlm` or `optim` optimizers or (ii) in SPlus, via `nlmin`. The $\hat{B}$ matrix in (4.9) is obtained by specifying the `hessian` option, available in both `nlm` and `nlminb`. The `gradient` option does not give the individual duration model scores $U_i(\hat{\theta}, \hat{\alpha})$, it only gives the value of the gradient evaluated at the maximum. Therefore, extra code would be needed to compute the terms $U_i(\hat{\theta}, \hat{\alpha})$. Other software with good optimization functions (e.g. the `proc nlp` procedure in SAS) could also be used to obtain estimates and hessian matrices.

Another way to obtain the maximum in (4.2) is by using the function `censorReg` in SPlus specifying the desired distribution $f_0(\cdot)$ in (4.15). The matrix $\hat{B}$ can be obtained from the variance matrix of the estimated coefficients given by the output and the $U_i(\hat{\theta}, \hat{\alpha})$ score residuals can be obtained with some extra code. A description on how to specify `censorReg` in order to deal with weights, right censoring and delayed entry can be found in Cook and Lawless, [15] (Appendix C).

Estimation of the loss to follow-up related probabilities $p_{it}(\hat{\alpha})$ can be done using `proc logistic` or `glm` in SAS and R/SPlus, respectively. When the loss to follow-

up scores $G_i(\hat{\alpha})$ cannot be obtained directly from the software (which is the case in SAS), it is useful to note that the score residuals are related to the DFBETA residuals, calculated as the approximate change $(\hat{\alpha} - \hat{\alpha}_i^1)$ in the vector of parameter estimates due to the omission of the ith observation. The relationship between the score residuals $G_i(\hat{\alpha})$ and the DFBETA residuals $\Delta\hat{\alpha}_i^1$ is given by:

$$\Delta\hat{\alpha}_i^1 \simeq \left(\frac{R_{it} - \lambda_{it}(\hat{\alpha}_t)}{1 - h_{ii}}\right) Z_i^c(t)' \; \hat{Var}(\hat{\alpha}) = \left(\frac{G_i(\hat{\alpha})}{1 - h_{ii}}\right) \hat{Var}(\hat{\alpha})$$

where $h_{ii} = Z_i^c(t) \; \hat{Var}(\hat{\alpha}) \; Z_i^c(t)'$ is the (i,i)th element of the "Hat" matrix based on the logistic regression and $\hat{Var}(\hat{\alpha})$ is the variance matrix of the estimated $\hat{\alpha}$ coefficient by the software (see Collet [14], chapter 5).

4.2 A simulation

This simulation is motivated by the observational framework used in the UK Millennium Cohort Study, in which children are followed longitudinally at ages of 9 months, and 3,5,7 and 9 years. This survey collects information about a variety of characteristics regarding children's growth, including features from their home environment, such as their family's health, economic status and composition. For simplicity, we consider simple random samples of individuals.

In this simulation, we will assume that the time to achieve certain milestones in the growth of children is of interest. Let's suppose that a sample of children is followed at ages 1,3,5,7, and 9. Denote the time to the event of interest by Y_i, experienced by individual i. It is simulated here as a Normal random variable with mean and variance given by

$$E(Y_i|x_{1i}, x_{2i}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \;\text{ and }\; Var(Y_i|x_{1i}, x_{2i}) = \sigma^2. \tag{4.16}$$

Further, suppose that the covariates in (4.16) follow a bivariate normal distribution, with mean $\mu = (\mu_{x1}, \mu_{x2})$, and variances $Var(X_1) = \sigma_{x1}^2$, $Var(X_2) = \sigma_{x2}^2$ and covariance $Cov(X_1, X_2) = \sigma_{x1,x2}^2$. The correlation between X_1 and X_2 is given by $\rho = Cov(X_1, X_2)/(\sigma_{x1} \cdot \sigma_{x2})$.

Further, suppose that each child is subject to loss to follow-up before the end of the study and that $\lambda_{it}(\alpha) = Pr(R_{it}|R_{i,t-1} = 1, x_{2i}; \alpha)$, that is, the probability of being observed at time t given that the person was also observed at $t-1$, depends on X_2. Loss to follow-up is simulated from the following logistic model:

$$\text{logit}\lambda_{it}(\alpha) = \alpha_0 + \alpha_1 x_{2i}, \quad \text{for } \ t = 1, 3, 5, 7, 9. \tag{4.17}$$

This model is also used to estimate the dropout related probabilities

$$p_{it}(\alpha) = Pr(R_{it}|x_{2i}; \alpha) = \lambda_i^t,$$

where $\alpha = (\alpha_0, \alpha_1)'$ and $\lambda_i = \lambda_{i1} = \lambda_{i2} = \cdots \lambda_{it}$. Note that in this case we have that $p_{it}(\alpha) = p_i(\alpha)$ for all t and the parameters to be estimated from this model are given by $\alpha = (\alpha_0, \alpha_1)'$. Note that in real life we would not have knowledge about the LTF process and should base the estimates of the IPC weights on separate LTF models for $t \in \{1, \ldots, 5\}$; however in this particular simulation we fitted only one model for computational convenience.

It is of interest to examine the behavior of the IPCW method in the presence of dependent loss to follow-up. The working model we use for the simulated durations Y_i is Normal with mean and variance given by

$$E(Y_i|x_{1i}) = \beta_0^w + \beta_1^w x_{1i} \ \text{ and } \ Var(Y_i|x_{1i}) = \sigma_w^2. \tag{4.18}$$

From the properties of the Normal distribution, the true values of the parameters in (4.18) are

$$\begin{aligned}
\beta_0^w &= \beta_0 + \beta_2\big(\mu_{x2} - \rho\ (\sigma_{x2}/\sigma_{x1})\mu_{x1}\big), \\
\beta_1^w &= \beta_1 + \beta_2 \rho\ \sigma_{x2}/\sigma_{x1}, \text{ and} \\
\sigma_w^2 &= \beta_2^2 \sigma_{x2}^2 (1 - \rho^2) + \sigma^2;
\end{aligned}$$

where $\beta_0, \beta_1, \beta_2, \rho$ and σ^2 are the parameters in (4.16). We will investigate how well a working model with (4.18) estimates these values.

When the variable X_2 affects both durations and dropout ($\beta_2 > 0$) and is not included in the working duration model (4.18), the IPC weights are necessary to

achieve consistent estimation of β_0^w, β_1^w and σ_w^2. It is of interest to show how the IPCW method behaves in different scenarios.

With Y_i in years; let the overall variation of Y_i be denoted by $Var\big(Y_i\big) = \sigma_y^2$. This variation has been set to be $\sigma_y^2 = 0.64$. Based on this and $\beta_0 = 6$, scenarios for simulation are considered with (i) proportions of explained variation of $EV = 1 - \sigma^2/\sigma_y^2 = 0.3$ and 0.5; (ii) correlation values between X_1 and X_2 of $\rho = 0$ and 0.4; and (iii) duration model coefficients $\beta_2 = 0$ and $\beta_1 = \beta_2$. The values of β_1 and β_2 can be obtained from EV and by noting that $\sigma_y^2 = \beta_1^2 + \beta_2^2 + 2\rho\beta_1\beta_2 + \sigma^2$ when choosing, without loss of generality, $\mu_{x1} = \mu_{x2} = 0$ and $\sigma_{x1}^2 = \sigma_{x2}^2 = 1$. The values defined in (ii) and (iii) give eight possible scenarios, found in Table 4.1.

We employ an initial sample size of 1450. Thus, one repetition of the simulation consists of creating a set of $n = 1450$ independent durations $Y_1, Y_2 \ldots Y_n$ from model (4.16) and then simulating a loss to follow-up time $t \in \{1, 3, 5, 7, 9\}$ in years for each, based on the model in (4.17). The values of the parameters in the dropout model have been set to achieve approximately 50% loss to follow-up by the end of the study (by $t = 9$) and by considering $Pr(R_{i9} = 1|x_{2i} = -1.645) = .75$ and $Pr(R_{i9} = 1|x_{2i} = 1.645) = .25$. These probabilities give $\alpha_0 = 1.984$ and $\alpha_1 = -.5123$. We simulated 1000 independent samples for each scenario.

The total of 1000 simulations within each scenario gave an average of 1260 observed spells from which about 33% were censored. Note that the number of spells differs from the initial sample size of 1450 because some individuals were lost to follow-up in the first year (see (4.17)) and so no data on them were collected. The average proportion of individuals lost to follow-up by year 9 was 0.48. The proportion of censored spells and individuals lost to follow-up by year 9 did not vary substantially across scenarios; the individual values can be found in Table 4.2.

Results from the simulations for each of the four possible scenarios for the explained variation factors of $EV = 0.3$ and $EV = 0.5$ can be found in Tables 4.3-4.4 and 4.5-4.6, respectively. From these Tables, it is possible to observe the behavior of unweighted estimates and estimates based on the IPC weights. The column de-

Table 4.1: Scenarios used for simulation of durations based on the Millennium Cohort Study's framework. Regression parameters with Y in both years and months are shown.

			Year Scale		Month Scale	
Scenario	EV	ρ	β_2	β_1	β_2	β_1
1	0.3	0.0000	0.3098	0.3098	3.7181	3.7181
2		0.0000	0.0000	0.4382	0.0000	5.2581
3		0.4000	0.2619	0.2619	3.1428	3.1428
4		0.4000	0.0000	0.4382	0.0000	5.2581
5	0.5	0.0000	0.4000	0.4000	4.8000	4.8000
6		0.0000	0.0000	0.5657	0.0000	6.7882
7		0.4000	0.3381	0.3381	4.0572	4.0572
8		0.4000	0.0000	0.5657	0.0000	6.7882

Table 4.2: Average proportion of censored spells, number of observed spells and average proportion of LTF individuals by year 9, across 1000 samples.

Scenario	Prop.Censored	Av.No.Spells	Prop.LTF
1	0.328	1259	0.485
2	0.326	1259	0.484
3	0.328	1260	0.484
4	0.327	1260	0.484
5	0.329	1260	0.485
6	0.326	1260	0.484
7	0.330	1260	0.485
8	0.327	1260	0.483

noted by "True" represents the values of β_0^w, β_1^w and σ_w under the true model. For each one of these parameters, the column labeled as "Av.Coef" gives the average of the estimated coefficients over the 1000 samples, "SD.Coef" gives the standard error of the estimated coefficients, "Av.Bias" is the average of the bias, "Av.SD" is the average standard deviation calculated under the unweighted method (where "Method=Unw") and the IPC method ("Method=IPC"). The "Cov" column shows the estimated coverage probability obtained for each parameter, based on a nominal 95% confidence interval (Estimate $\pm$1.96 SD).

In the case of scenario 1 when $\rho = 0, \beta_2 = \beta_1$ and $EV = 0.3$ (Table 4.3), results for the Unw method give good coverage for β_1^w, with the IPC method giving a better result in terms of bias. The absence of the variable X_2 in the fitted model is reflected by a low coverage for β_0^w and $log(\sigma_w)$. The average standard error "Av.SD" is slightly higher for method IPC compared to Unw. This behavior appears consistent when increasing the explained variation to $EV = 0.5$ (scenario 5, Table 4.5), with the coverage for the slope parameter below the nominal value of 95%. Since X_2 affects both durations and dropout, IPC weights would be needed in this scenario. The IPC method shows good coverage and low bias for all parameters, with slight improvement at the higher EV value.

In scenarios 2 and 6 (Tables 4.3, 4.5), with $\rho = \beta_2 = 0$ and $EV = 0.3, 0.5$, IPC weights are not needed since X_2 does not affect dropout. Both Unw and IPC methods work well in these cases, with slightly better results when EV increases.

When $\rho \neq 0$ and $\beta_1 = \beta_2$, the variable X_2 affects both dropout and durations, hence IPC weights are expected to give better results than the Unw method. This is the case in terms of bias and coverage, as can be seen in Tables 4.4 and 4.6 (scenarios 3 and 7), where the IPC gives good coverage while the Unw gives coverage values below the nominal 95%. The behavior of the Unw method becomes worse as the EV increases to 0.5, while the IPC method shows consistent results for both values of EV.

Finally, when $\rho \neq 0$ and $\beta_2 = 0$, both Unw and IPC methods give good results

(scenarios 4 and 8 in Tables 4.4, and 4.6). The Unw method performs well because Y is independent of X_2 given X_1, and therefore there is no need for IPC weights.

This simulation shows that the IPC method performs better than Unw under dependent loss to follow-up, when no sampling design features are present. It is of interest to further explore the performance of the IPC method in hypothetical situations where the sampling mechanism has an effect. This will be analyzed in chapters 5 and 6, where Kaplan-Meier estimates and Cox PH models are implemented using the variance estimation techniques for survey data described in the following section.

Table 4.3: Results from simulation based on MCS study. Scenarios 1 and 2: $EV = 0.3, \rho = 0$ and $\beta_1 = \beta_2 = 3.7181$; $\beta_1 = 5.2581, \beta_2 = 0$ (month scale).

Scenario	Parameter	Method	True	Av.Coef	SD.Coef	Av.Bias	Av.SD	Cov
1	β_0^w	Unw	72.0000	71.1812	0.2986	-0.8188	0.2937	0.201
	β_1^w	Unw	3.7181	3.6889	0.2855	-0.0292	0.2935	0.953
	$\log(\sigma_w)$	Unw	2.1805	2.1655	0.0235	-0.0150	0.0239	0.901
	β_0^w	IPC	72.0000	72.0062	0.3050	0.0062	0.3007	0.943
	β_1^w	IPC	3.7181	3.7256	0.3067	0.0075	0.3150	0.955
	$\log(\sigma_w)$	IPC	2.1805	2.1792	0.0266	-0.0013	0.0263	0.950
2	β_0^w	Unw	72.0000	71.9964	0.2782	-0.0036	0.2709	0.944
	β_1^w	Unw	5.2581	5.2387	0.2745	-0.0194	0.2707	0.946
	log(scale)	Unw	2.0834	2.0821	0.0243	-0.0013	0.0239	0.956
	β_0^w	IPC	72.0000	71.9951	0.2898	-0.0049	0.2809	0.943
	β_1^w	IPC	5.2581	5.2384	0.2848	-0.0197	0.2803	0.949
	$\log(\sigma_w)$	IPC	2.0834	2.0821	0.0255	-0.0013	0.0248	0.949

Table 4.4: Results from simulation based on MCS study. Scenarios 3 and 4: $EV = .3, \rho = .4$ and $\beta_1 = \beta_2 = 3.1428$; $\beta_1 = 5.2581, \beta_2 = 0$ (month scale).

Scenario	Parameter	Method	True	Av.Coef	SD.Coef	Av.Bias	Av.SD	Cov
3	Intercept	Unw	72.0000	71.3847	0.2889	-0.6153	0.2867	0.429
	x1	Unw	4.3993	4.2688	0.2868	-0.1305	0.2878	0.928
	log(scale)	Unw	2.1439	2.1337	0.0235	-0.0102	0.0239	0.929
	Intercept	IPC	72.0000	71.9798	0.2934	-0.0202	0.2945	0.948
	x1	IPC	4.3993	4.3875	0.3099	-0.0118	0.3076	0.949
	log(scale)	IPC	2.1439	2.1417	0.0251	-0.0022	0.0257	0.951
4	Intercept	Unw	72.0000	71.9843	0.2683	-0.0157	0.2729	0.957
	x1	Unw	5.2581	5.2650	0.2694	0.0068	0.2738	0.958
	log(scale)	Unw	2.0834	2.0820	0.0242	-0.0014	0.0239	0.949
	Intercept	IPC	72.0000	71.9826	0.2784	-0.0174	0.2822	0.949
	x1	IPC	5.2581	5.2621	0.2843	0.0039	0.2884	0.958
	log(scale)	IPC	2.0834	2.0819	0.0253	-0.0016	0.0249	0.946

Table 4.5: Results from simulation based on MCS study. Scenarios 5 and 6: $EV = .5, \rho = 0$ and $\beta_1 = \beta_2 = 4.8$; $\beta_1 = 6.7882, \beta_2 = 0$ (month scale).

Scenario	Parameter	Method	True	Av.Coef	SD.Coef	Av.Bias	Av.SD	Cov
5	Intercept	Unw	72.0000	70.9368	0.2631	-1.0632	0.2738	0.025
	x1	Unw	4.8000	4.7422	0.2828	-0.0578	0.2735	0.934
	log(scale)	Unw	2.1179	2.0937	0.0238	-0.0242	0.0240	0.832
	Intercept	IPC	72.0000	72.0076	0.2749	0.0076	0.2736	0.949
	x1	IPC	4.8000	4.7949	0.3212	-0.0051	0.3021	0.940
	log(scale)	IPC	2.1179	2.1164	0.0280	-0.0016	0.0271	0.939
6	Intercept	Unw	72.0000	72.0045	0.2236	0.0045	0.2294	0.952
	x1	Unw	6.7882	6.7843	0.2303	-0.0039	0.2288	0.951
	log(scale)	Unw	1.9152	1.9131	0.0240	-0.0020	0.0240	0.959
	Intercept	IPC	72.0000	72.0035	0.2341	0.0035	0.2376	0.954
	x1	IPC	6.7882	6.7836	0.2403	-0.0046	0.2370	0.952
	log(scale)	IPC	1.9152	1.9127	0.0248	-0.0025	0.0248	0.955

Table 4.6: Results from simulation based on MCS study. Scenarios 7 and 8: $EV = .5, \rho = .4$ and $\beta_1 = \beta_2 = 4.0572$; $\beta_1 = 6.7882, \beta_2 = 0$ (month scale).

Scenario	Parameter	Method	True	Av.Coef	SD.Coef	Av.Bias	Av.SD	Cov
7	Intercept	Unw	72.0000	71.2095	0.2663	-0.7905	0.2593	0.147
	x1	Unw	5.6794	5.5167	0.2604	-0.1627	0.2605	0.902
	log(scale)	Unw	2.0464	2.0300	0.0254	-0.0164	0.0240	0.873
	Intercept	IPC	72.0000	71.9892	0.2727	-0.0108	0.2631	0.942
	x1	IPC	5.6794	5.6811	0.2930	0.0017	0.2850	0.944
	log(scale)	IPC	2.0464	2.0436	0.0286	-0.0028	0.0264	0.926
8	Intercept	Unw	72.0000	72.0012	0.2278	0.0012	0.2314	0.956
	x1	Unw	6.7882	6.7908	0.2278	0.0026	0.2322	0.960
	log(scale)	Unw	1.9152	1.9133	0.0232	-0.0019	0.0240	0.948
	Intercept	IPC	72.0000	72.0036	0.2358	0.0036	0.2395	0.955
	x1	IPC	6.7882	6.7881	0.2431	-0.0001	0.2458	0.955
	log(scale)	IPC	1.9152	1.9132	0.0248	-0.0020	0.0250	0.944

4.3 Estimation from survey data

In the context of survey data, suppose that a sampling design was used to obtain a sample from a finite population of size N. Suppose that the sample $S = \bigcup_{r=1}^{R} \bigcup_{k=1}^{K_r} S_{rk}$ is composed of K_r clusters within R strata, $r = 1, ..., R$, where S_{rk} is the subsample corresponding to the kth cluster within the rth stratum. Let $I_i = I(i \in S)$ indicate whether individual i was included in the sample and the sampling probabilities be $\pi_i = Pr(I_i = 1)$, where π_i depends on the stratum i is in. The system of equations analogous to the ones in expression (4.1) are:

$$U(\theta, \alpha) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i \in S_{rk}} U_i(\theta, \alpha) = 0 \tag{4.19}$$

$$G(\alpha) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i \in S_{rk}} G_i(\alpha) = 0. \tag{4.20}$$

The individual estimating equations $U_i(\theta, \alpha)$ referring to the duration model

have a similar form as in (4.2),

$$U_i(\theta, \alpha) = \sum_{j=1}^{m_i} \sum_{t=1}^{M} w_{it}(y) U_{ijt}(\theta), \tag{4.21}$$

where $w_{it}(\alpha) = R_{it}/\pi_i p_{it}(\alpha)$, and $U_{ijt}(\theta) = \partial l_{ijt}(\theta)/\partial\theta$. The estimating equations referring to the model for loss to follow-up have the same form as in (4.4). As before, estimates of (θ, α) are obtained as solutions of these equations.

The discussion in Chapter 3 shows how using the IPCW method achieves unbiasedness under certain assumptions, in a general context. In particular, in order to show that (4.2) is unbiased for every t, let $H_i(M) = \{D_i(1), \ldots D_i(M)\}$ represent the duration history of individual i in the population, over the observation period $(0, M]$. Let Z_i^D denote a variable that contains information about the sample design, that is, about stratification and clustering. Let $Z_i(t)$ the set of explanatory variables of the duration model, which includes external covariates and history $H_i(t-1)$ up to time $t-1$. Further, consider $Z_i^c(t)$, the set of explanatory variables in the model for loss to follow-up in (4.5), which may include $Z_i(t)$.

As before, let us assume that (i) R_{it} is conditionally independent of $D_i(t)$ given $(Z_i^c(t), Z_i^D)$ (missing at random assumption, Robins et al. [51]), (ii) I_i is conditionally independent of $(D_i(t), Z_i^c(t))$ given Z_i^D, and (iii) the duration model is correctly specified, so that $E(\partial l_{it}(\theta)/\partial\theta | Z_i(t)) = 0$. Under these assumptions, we have that $U_{it} = U_{it}(\theta, \alpha)$ is unbiased for every t, $t = 1, \ldots, M$. The expected value of

$$U_{it} = \frac{R_{it}}{\pi_i p_{it}(\alpha)} \sum_{j=1}^{m_i} \frac{\partial l_{ijt}(\theta)}{\partial\theta}, \tag{4.22}$$

conditional on $Z^c(t), Z^D$, and $Z(t)$ is given by:

$$\begin{aligned} E_{D(t),I,R_t,Z^c(t)|Z^D,Z_i(t)}\{U_{it}\} &= E_{D(t),Z^c(t)|Z^D,Z_i(t)}\left\{E_{R_t,I|D(t),Z^c(t),Z^D,Z_i(t)}\{U_{it}\}\right\} \\ &= E_{D(t)|Z^D,Z_i(t)}\left\{E_{R_t|Z^c(t),Z^D} E_{I|Z^D}\{U_{it}\}\right\} \end{aligned} \tag{4.23}$$

by assumptions (i) and (ii). After applying these two expectations to U_{it}, and noting that $E_{D(t),Z^c(t)|Z^D,Z(t)}\{U_{it}\} = E_{D(t)|Z^D,Z(t)}\{U_{it}\}$ (see expression (3.7)), we

are left with:

$$E_{D(t)|Z^D,Z_i(t)}\{U_{it}\} = E_{D(t)|Z^D,Z_i(t)}\left\{\sum_{j=1}^{m_i}\partial l_{ijt}(\theta)/\partial\theta\right\} \tag{4.24}$$

$$= 0, \text{ by assumption } (iii). \tag{4.25}$$

Variance estimates

Miller et al. [44] adapted the variance estimation procedure with IPC weights in Robins et al. [51] to the context of survey data. The variance for $\hat{\theta}$ has the same "sandwich" form as in (4.9). Let it be denoted as the combined IPCW and design-variance estimate for $\hat{\theta}$, given by

$$\widehat{Var}(\hat{\theta})_{comb} = \hat{B}_{comb}^{-1}\hat{C}_{comb}\hat{B}_{comb}^{-1\,\prime}, \tag{4.26}$$

where

$$\hat{B}_{comb} = -\sum_{r=1}^{R}\sum_{k=1}^{K_r}\sum_{i\in S_{rk}}\frac{\partial U_i(\theta,\hat{\alpha})}{\partial\theta}\Big|_{\hat{\theta}} \quad \text{and} \tag{4.27}$$

$$\hat{C}_{comb} = \sum_{r=1}^{R}\frac{K_r}{K_r-1}\sum_{k=1}^{K_r}\left\{\left(\sum_{i\in S_{rk}}\hat{E}_i\right) - \left(\frac{1}{K_r}\sum_{k=1}^{K_r}\sum_{i\in S_{rk}}\hat{E}_i\right)\right\}^{\otimes 2} \tag{4.28}$$

where

$$\hat{E}_i = U_i(\hat{\theta},\hat{\alpha}) - \left[\sum_{r=1}^{R}\sum_{k=1}^{K_r}\sum_{i\in S_{rk}}U_i(\hat{\theta},\hat{\alpha})G_i(\hat{\alpha})'\right]\left[\sum_{r=1}^{R}\sum_{k=1}^{K_r}\sum_{i\in S_{rk}}G_i(\hat{\alpha})G_i(\hat{\alpha})'\right]^{-1}G_i(\hat{\alpha}),$$

where $U_i(\theta,\alpha)$ is as in (4.21) and $A^{\otimes 2} = AA'$. Miller et al. obtained the terms $\hat{E}_i$ above in the $\hat{C}$ matrix from Robins et al. in (4.11), and applied to their sum the standard stratified sampling variance estimator (4.28) to account for the stratified and cluster sampling aspects of the data (Cochran, [13]).

As discussed in section 4.1, the variance estimates with the IPC weights alone described in expression (4.2) are smaller than the variance estimates considering the IPC weights as fixed. It is of interest to examine the behavior of variance estimates that use combined IPC and sampling weights as in (4.26). This will be

explored in the remainder chapters of this thesis. However, we note that a variance estimate analogous to the form (4.8) can also be given in the survey context, thus suggesting that treating the IPC weights, and hence the full combined weights, as fixed produces a (slightly) larger variance estimate than treating them as random.

Chapter 5

Weighted Kaplan-Meier Estimation

As descriptive quantities, estimates of survivor distributions from a finite population are of interest, for example, the distribution of jobless spells for individuals living in different provinces. We later consider these based on labour data collected by the Survey of Labour and Income Dynamics (SLID). One particularity of this study would be that not every person in the population experiences the event of interest (being jobless) in a given period of time. In other settings, interest might lie on the survivor distribution of the time to an event, for instance, the study of the time by which children reach a certain developmental milestone, from data gathered from the UK Millenium Cohort Study (MCS). Note that in this example a high proportion of the population of interest, if not its totality, may experience the event in question.

Weighted Kaplan-Meier (K-M) estimates for dependent loss to follow-up in the non-survey setting have been considered by Robins and Finkelstein [50], Robins [49], Satten et al. [55]. Several authors have considered them in the survey data context, for example, Folsom et al. [20], and Korn and Graubard [28]; however, the issue of dependent loss to follow-up is not accounted for. Especially in the case of estimating a duration distribution without covariates, dependent loss to

follow-up becomes a problem, since it is typically related to covariates and previous event history variables that are not being accounted for in the estimation of the population survivor distribution.

The methods developed in chapter 4 concern parametric modelling as the basis for variance estimation in the presence of dependent loss to follow-up. It is possible to apply the methodology to Kaplan-Meier estimation, since, as we show below, it can be performed through the estimation of the discrete-time hazard function via parametric likelihood methods.

The first section of this chapter provides a framework in which the notion of a finite population distribution is introduced, as well as notation. Section 5.2 gives a discussion of estimation of the survivor distribution using IPCW methods, accounting for the randomness of the weights. This discussion makes reference to the methods in chapter 4.

Section 5.3 gives a simulation study on the performance of the IPC weights in the presence of dependent loss to follow-up and stratification effects. Results from using sampling design weights with and without IPC weights are compared.

5.1 Framework

Let $(0, M]$ denote the follow-up period for individuals $i = 1, 2, ..., n$ and consider the notation defined at the beginning of section 4.1 regarding sequences of durations. Briefly, for an individual i in the population, let $m_i \geq 0$ denote the number of durations that the individual experienced within the observation period. As before, denote u_{ij}, v_{ij} as the start and end times of the durations represented by $Y_{ij} = v_{ij} - u_{ij}$; $j = 1, \ldots m_i$ and the time a person was last seen as $C_i \in \{1, 2, ..., M\}$.

As discussed in section 2.3, the duration distribution as a finite population quantity is expressed as:

$$S_{\mathcal{U}}(y) = \frac{1}{N} \sum_{i \in \mathcal{U}} \sum_{j=1}^{m_i} I(Y_{ij} \geq y). \tag{5.1}$$

For individuals i with $m_i = 0$, the corresponding summand in (5.1) is equal to zero. In (5.1), $\mathcal{U}$ is the finite population of size N^* and $N = \sum_{i=1}^{N^*} m_i$ is the total number of durations in the population.

For the discussion that follows, it is useful to assume that the finite population quantity in (5.1) converges in probability to a superpopulation duration distribution $S(y)$, as the population's size N^* increases to infinity. This is reasonable since the durations Y_{ij} and N are latent random variables at the time the sample is selected, and in this sense, the finite population quantity in (5.1) has random components. This is expressed as follows,

$$S(y) = plim\ S_{\mathcal{U}}(y) = plim\ \left(\frac{N^*}{N}\right) plim\ \left(\frac{1}{N^*}\sum_{i\in\mathcal{U}}\sum_{j=1}^{m_i} I(Y_{ij} \geq y)\right).$$

In large-scale population surveys, it might be of more interest to study a duration survivor distribution for a particular stratum or group in the population (e.g., a specific province in Canada), rather than an overall estimate that combines all strata. The reasoning used above can be easily applied to this case.

5.2 Point and variance estimation

Let $I_i = I(i \in S)$ indicate whether individual i was included in the sample and let the sample inclusion probability be represented by $\pi_i = Pr(I_i = 1|Z_i^D)$, where Z_i^D is a set of factors regarding the sampling design. Let's consider a discrete time scale and duration times without including clusters and strata for now. From section 1.2, we have the hazard function denoted by $h(y) = \Pr(Y = y|Y \geq y) = f(y)/S(y)$, where $S(y)$ and $f(y) = S(y) - S(y+1)$ (when $S(y) = P(Y \geq y)$ are the corresponding discrete survivor and probability functions, respectively.

Let T be an upper limit on the duration variable Y and define $\theta = (h(1), ..., h(T))'$, where $h(y)$ is the hazard function for $y = 1, ..., T$. For individuals that had at least one event in the observation period, that is, those who had $m_i \geq 1$, the estimating

functions $U(\theta, \alpha)$ in (4.2) have elements of the form:

$$U_i(\theta,\alpha)_y = \sum_{t=1}^{M}\sum_{j=1}^{m_i} \frac{\delta_{ijt}(y)}{\pi_i p_{it}(\alpha)} \left[d_{ijt}(y) - h(y)\right], \\ = \sum_{t=1}^{M} U_{it}(\theta,\alpha)_y \qquad\qquad y = 1,\ldots,T \tag{5.2}$$

where

$$d_{ijt}(y) = I\left(Y_{ij} = y, t-1 < u_{ij} + y \leq t\right) R_{it}, \tag{5.3}$$

$$\delta_{ijt}(y) = I\left(Y_{ij} \geq y, t-1 < u_{ij} + y \leq t\right) R_{it}; \tag{5.4}$$

u_{ij} is the start time of the j-th duration Y_{ij}, $\pi_i = Pr(i \in S | Z_i^D)$ is the sample inclusion probability and $p_{it}(\alpha) = \lambda_{i1}(\alpha)\ldots\lambda_{it}(\alpha)$ the probability of being observed at time t, estimated from the estimating equations in (4.4) and the logistic model in (4.5).

The terms in (5.2) are unbiased for every t. Recall that it is assumed that the covariate vectors $Z_i^C(t)$ in the loss to follow-up modelling (expressions (4.4) and (4.5)) include enough information such that R_{it} is conditionally independent of $D_i(t)$, given $Z_i^C(t)$. The variables in $Z_i^C(t)$ must include terms that affect both loss to follow-up and durations. It is important to note that $Z_i(t)$ may depend on covariates or previous event information only up to time $t-1$. This ensures that data at t when the person is lost to follow-up, and later, are missing at random. That is, that R_{it} is independent of the full duration history $H_i(M)$ given $Z_i^C(t)$. In line with the discussion in 4.1.2, analogous to (4.24) we have that,

$$E_{H(M)R_t,R|Z^C(t),Z^D}\left\{U_{it}\right\} = E_{H(M)|Z^C(t),Z^D}\left\{\sum_{j=1}^{m_i} I(Y_{ij} \geq y)\left[I(Y_{ij} = y) - h(y)\right]\right\} \tag{5.5}$$

$$= \left\{m_i f(y) - m_i S(y) h(y)\right\} \tag{5.6}$$

$$= 0. \tag{5.7}$$

The solution to $\sum_{i=1}^{n} U_i(\theta, \hat{\alpha}) = 0$ in (5.2) is:

$$\hat{h}(y) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i} \hat{w}_{ij}(y) d_{ij}(y)}{\sum_{i=1}^{n}\sum_{j=1}^{m_i} \hat{w}_{ij}(y)}, \tag{5.8}$$

where $d_{ij}(y) = \sum_{t=1}^{M} d_{ijt}(y)$, $\hat{w_{ij}}(y) = \sum_{t=1}^{M} \hat{w}_{ijt}(y)$, $\hat{w}_{ijt}(y) = \frac{\delta_{ijt}(y)}{\pi_i p_{it}(\hat{\alpha})}$, and $\hat{\alpha}$ is a consistent estimator of α obtained from the estimating functions corresponding to the dropout model as in (4.4) and (4.5). The weighted KM estimator of $S(y)$ is then given by

$$\hat{S}(y) = \prod_{s=1}^{y-1} \Big(1 - \hat{h}(s)\Big), \qquad y = 1, 2, \ldots, T. \tag{5.9}$$

Variance estimation

Consider that a sample $S = \bigcup_{r=1}^{R} \bigcup_{k=1}^{K_r} S_{rk}$ of size n, based on K_r clusters within strata $r = 1, ..., R$. The system of estimating equations can be expressed as in (4.19) and (4.20), with a corresponding change in (5.8).

Let $\widehat{Var}_{comb}(\hat{\theta}) = \hat{B}_{comb}^{-1} \hat{C}_{comb} \hat{B}_{comb}^{-1\,\prime}$ denote the asymptotic variance of the estimated parameter $\hat{\theta} = (\hat{h}(1), \ldots, \hat{h}(T))'$ when using the sampling design weights combined with the IPCW, as in (4.26). For the KM estimation, the B_{comb} matrix in (4.27) is a diagonal matrix with terms based on:

$$\left[\frac{\partial\, U_i(\theta, \hat{\alpha})}{\partial \theta}\Big|_{\hat{\theta}}\right]_y = \sum_{t=1}^{M} \sum_{j=1}^{m_i} \frac{1}{\pi_i p_{it}(\hat{\alpha})} \left\{ -\frac{d_{ijt}(y)}{\hat{h}(y)^2} - \frac{\delta_{ijt}(y) - d_{ijt}(y)}{[1 - \hat{h}(y)]^2} \right\}; \tag{5.10}$$

where $U_i(\theta, \hat{\alpha}) = \big(U_i(h(1), \hat{\alpha}), ..., U_i(h(T), \hat{\alpha})\big)'$ are the combined IPC and design weighted score residuals in (5.2). The middle matrix C_{comb} is obtained as in (4.28), using $U_i(\theta, \hat{\alpha})$ as in (5.2).

An asymptotic variance estimate for $\hat{S}(y)$ in (5.9) is given by a straightforward application of the delta theorem ([32], Appendix B.1), leading to

$$\widehat{Var}\left\{\hat{S}(y)\right\} = \hat{S}(y)^2 \sum_{s=1}^{y-1} \sum_{t=1}^{y-1} \frac{\widehat{Cov}\left[\hat{h}(s), \hat{h}(t)\right]}{\left[1 - \hat{h}(s)\right]\left[1 - \hat{h}(t)\right]}, \tag{5.11}$$

where $\widehat{Cov}[\hat{h}(s), \hat{h}(t)]$ is the (s, t) element of (4.26) with $\hat{B}_{comb}$ and $\hat{C}_{comb}$ as in (4.27) and (4.28), respectively.

5.3 A simulation study

The objective of this simulation is to assess estimates (5.9) of the survivor function in (5.1) and their estimated variance (5.11), based on multiple spells from individuals sampled from a stratified finite population. The main interest is to show the performance of the weighted Kaplan-Meier method using two choices of weights, mainly, design weights π_i^{-1} and a combination of the design and IPC weights $\pi_{it}^{-1} = (\pi_i \hat{p}_{it})^{-1}$.

5.3.1 Setup

The simulated process is motivated by that of jobless spells from SLID. It consists of an alternation of durations of sojourns in the states "jobless" and "not jobless", termed as NJ and J, similar to the example of alternating between employment and unemployment in section 3.3, depicted in Figure 3.1. For convenience, we assume that the process started in state NJ at $t = 0$. We will denote the durations in state J as Y_{ij} and those in state NJ as Y_{ij}^{NJ} for individual i, so that a whole sequence is labeled as $Y_{i1}^{NJ}, Y_{i1}, Y_{i2}^{NJ}, Y_{i2}, \ldots$. Interest resides in estimation of the distribution of jobless spells over some time period.

The sequences of durations in states $\{NJ, J\}$ are generated for a finite population of individuals, of size N. This population $\mathcal{U}$ is composed of ten strata $\mathcal{U}_1, \ldots \mathcal{U}_{10}$, and within each stratum, a simple random sample of individuals is obtained.

Individuals are simulated to have sequences of jobless and not jobless spells over a period of six years (312 weeks), and the durations of the spells are measured in weeks. Every individual will start with an NJ spell with duration Y_{i1}^{NJ} of a certain length, which will be followed by a J spell with duration Y_{i1}, which in turn is followed by a second NJ duration Y_{i2}^{NJ}, and so on. Simulation of individual processes stop when the sum of the sequence of durations is greater or equal to 312 weeks.

In the population, the multiple jobless spell duration times Y_{ijr} are generated independently from a log-Normal model where $Y^*_{ijr} = \log(Y_{ijr})$; $i = 1, \ldots n$ and $j = 1, \ldots, m_i$, $r = 1, ..., 10$ and $m_i \geq 0$. The distribution of $[Y^*|X_1, X_2, \alpha^*]$ is given by:

$$E(Y^*_{ijr}|x_{ir1}, x_{ir2}, \alpha^*_r) = \beta_{0r} + \beta_1 x_{ir1} + \beta_2 x_{ir2} + \epsilon_{ijr}, \tag{5.12}$$

$$\beta_{0r} = \beta_0 + \alpha^*_r, \quad \epsilon_{ijr} \sim N(0, \sigma^2);$$

where $(X_1, X_2) \sim$ Bivariate Normal with vector mean $\mu = (\mu_{x_1}, \mu_{x_2})$ and a variance matrix with elements $Var(X_1) = \sigma^2_{x_1}$, $Var(X_2) = \sigma^2_{x_2}$ and covariance $\text{Cov}(X_1, X_2) = \sigma_{x_1,x_2}$. The intercept is further defined in terms of $\alpha^* = (\alpha^*_1, \ldots, \alpha^*_{10})' =$ (-0.366, -0.125, -0.119, -0.116, -0.055, -0.029, -0.006, 0.164, 0.318, 0.334)′, which is a vector of fixed stratum-specific effects. These were simulated from a Normal model with zero mean and variance equal to .084, and then centered about their mean. The variance of the generated fixed effects is $Var\big(\alpha^*\big) = 0.047$. The "not jobless" spells denoted as Y^{NJ}_{ij} are generated independently from an exponential distribution with mean $\gamma_1 \exp(\gamma_2 x_{i2})$.

Selection probabilities within each stratum in the population are specified so that higher sampling weights are assigned to strata with longer average durations, and are given by $p =$(0.02,0.022, ... ,0.038). Each stratum has 10000 individuals; then the sample sizes from the strata $r = 1, 2, \ldots, 10$ are (100,110,120, ... , 190), respectively, for a total sample size of 1450. No clustering is assumed and the simulated processes within strata are mutually independent.

For individuals that are selected in the sample, the loss to follow-up process is simulated from a logistic model, where

$$\text{logit}\lambda_{it}(\alpha) = \alpha_0 + \alpha_1 x_{i2} \quad t = 1, \ldots, 5. \tag{5.13}$$

with $\lambda_{it}(\alpha) = \Pr(R_{it}|R_{i,t-1} = 1, X_{i2} = x_{i2})$ and from where the IPC related probabilities are obtained: $p_{it}(\alpha) = \Pr(R_{it=1}|x_{i2}) = \lambda_{i1}\lambda_{i2} \cdots \lambda_{it}$.

The way in which this simulation is set up allows people in strata with larger indices to have longer jobless spells, higher probabilities of selection and higher

probabilities of becoming lost to follow-up, the latter by setting $\alpha_1 < 0$. It is assumed that the time a person is last seen, C_i, is conditionally independent of (Y_{ij}, X_{i1}) given x_{i2}.

In this simulation we have that the variable X_2 affects both durations and dropout. Since we are not including it in the estimation of the duration distribution, we expect to see that IPC weights are needed to achieve unbiasedness of the KM estimate.

5.3.2 Estimation formulas

The estimation formulas that were used for each weighting method are given in the following paragraphs. To keep the notation simple, it will be assumed that the terms $\hat{h}(y)$, $\hat{S}(y)$, $U_i(h(y), \alpha)$ correspond to the design, or combined weighted cases, according to the context in which they are mentioned below.

Design method:

The sampling probabilities are $\pi_i = Pr(i \in S_r) = n_r/N_r$, for individual i in stratum r, where S_r is the stratum r sample of size n_r and N_r is the size of the corresponding sub-population, $r = 1, 2, ..., R$. The estimate of the hazard function is given by (5.8), with $\hat{w}_{ij}(y) = \pi_i^{-1}$:

$$\hat{h}(y) = \frac{\sum_{r=1}^{R}(N_r/n_r)\sum_{i\in S_r}\sum_{j=1}^{m_i} I(Y_{ij} = y, \delta_i = 1)}{\sum_{r=1}^{R}(N_r/n_r)\sum_{i\in S_r}\sum_{j=1}^{m_i} I(Y_{ij} \geq y)}. \tag{5.14}$$

The superpopulation variance (based on Boudreau and Lawless, [9]) for the estimated survivor function $\widehat{Var}\big(\hat{S}(y)\big)$ has the same form as in (5.11), where $\widehat{Cov}\big(\hat{h}(s), \hat{h}(t)\big) = \hat{B}_{des}^{-1}\hat{C}_{sup}\hat{B}_{des}^{-1}$ and where $\hat{B}_{des}$ is a diagonal matrix of the form

$$\hat{B}_{des} = -\text{diag}\left\{\sum_{r=1}^{R}\sum_{i=1}^{n_r}\frac{\partial U_{ir}(h(y))}{\partial h(y)}\Big|_{\hat{h}(y)};\quad y = 1, \ldots, T\right\} \tag{5.15}$$

where

$$\frac{\partial\, U_{ir}(h(y))}{\partial\, h(y)}\Big|_{\hat{h}(y)} = \frac{N_r}{n_r}\sum_{j=1}^{m_i}\left\{-\frac{d_{ijt}(y)}{\hat{h}(y)^2} - \frac{\delta_{ijt}(y) - d_{ijt}(y)}{[1-\hat{h}(y)]^2}\right\}; \tag{5.16}$$

and where $\hat{h}(y)$ is the design estimated hazard function in (5.14). The middle matrix $\hat{C}_{sup}$ is of the form

$$\hat{C}_{sup} = \sum_{r=1}^{R} \sum_{i=1}^{n_r} U_{ir}(\hat{h}) U_{ir}(\hat{h})',$$

where $U_{ir}(\hat{h}) = \big(U_{ir}(\hat{h}(1)), ..., U_{ir}(\hat{h}(T))\big)'$ are the score residuals from individual i in stratum r as in (5.2), with $\pi_i = n_r/N_r$ and $p_{it}(\alpha)$=1.

The finite population variance (based on Binder, [4]) of $\hat{S}(y)$ is slightly different. It has the form in (5.11), where $\widehat{Cov}[\hat{h}(s), \hat{h}(t)] = \hat{B}_{des}^{-1} \hat{C}_{\text{fin}} \hat{B}_{des}^{-1}$, and

$$\hat{C}_{fin} = \sum_{r=1}^{R} \frac{n_r}{n_r - 1} \sum_{i=1}^{n_r} \Big(U_{ir}(\hat{h}) - \bar{U}_r(\hat{h}) \Big)^{\otimes 2}, \; A^{\otimes 2} = AA', \tag{5.17}$$

where $\bar{U}_r(\hat{h}) = \sum_{i=1}^{n_r} U_{ir}(\hat{h})/n_r$ and where $U_{ir}(\hat{h})$ and $\hat{h}$ are weighted with design weights.

Combined method:

This method goes along the lines of section 5.2, where the hazard function is estimated as in (5.8), where $\pi_i = n_r/N_r$, as before, is the probability of inclusion of individual i in stratum r. The survival function is estimated via (5.9) and the form of the variance estimate of $\widehat{S}(y)$ is (5.11), with $\widehat{Cov}[\hat{h}(s), \hat{h}(t)]$ as the (s,t) element of $\hat{B}_{comb}^{-1} \hat{C}_{comb} \hat{B}_{comb}^{-1}$, where the diagonal matrix $\hat{B}_{comb}^{-1}$ has elements of the form

$$\hat{B}_{comb} = -\text{diag} \left\{ \sum_{r=1}^{R} \sum_{i=1}^{n_r} \frac{\partial U_{ir}(h(y))}{\partial \; h(y)} \Big|_{\hat{h}(y)}; \quad y = 1, \ldots, T \right\} \tag{5.18}$$

and the middle $\hat{C}_{comb}$ matrix is:

$$\hat{C}_{comb} = \sum_{r=1}^{R} \frac{n_r}{n_r - 1} \sum_{i=1}^{n_r} \Big\{ E_{ir} - \bar{E}_r \Big\}^{\otimes 2},$$

$$E_{ir} = U_{ir} - (\sum_{i=1}^{n_r} U_{ir} G_{ir}')(\sum_{i=1}^{n_r} G_{ir} G_{ir}')^{-1} G_{ir},$$

where $A^{\otimes 2} = AA'$, $U_{ir} = \big(U_{ir}(\hat{h}(1), \hat{\alpha}), \ldots, U_{ir}(\hat{h}(T), \hat{\alpha})'$ are the combined IPC and design weighted score residuals from the duration model, $G_{ir} = G_{ir}(\hat{\alpha})$ the score residuals from the LTF model for individual i in stratum r, respectively; and $\bar{E}_r = \sum_{i=1}^{n_r} E_{ir}/n_r$ is the mean of E_{ir} from stratum r.

The **"naive"** variance of $\hat{S}(y)$ from the COMB method treats the IPC weights as fixed instead of random. It has the form in (5.11) and the covariance matrix for the hazard functions is given by $\hat{B}_{comb}^{-1}\hat{C}_{\text{N.comb}}\hat{B}_{comb}^{-1}$, where

$$\hat{C}_{N.comb} = \sum_{r=1}^{R} \frac{n_r}{n_r - 1} \sum_{i=1}^{n_r} \Big\{ U_{ir}(\hat{h}) - \bar{U}_r(\hat{h}) \Big\}^{\otimes 2}, \; A^{\otimes 2} = AA', \tag{5.19}$$

and where $\bar{U}_r(\hat{h}) = \sum_{i=1}^{n_r} U_{ir}(\hat{h})/n_r$ and where $U_{ir}(\hat{h}) = \big(U_{ir}(\hat{h}(1)), ..., U_{ir}(\hat{h}(T))\big)'$ and $\hat{h}$ are weighted with combined IPC and design weights.

5.3.3 Results

The parameter values were specified using a scheme similar to the one used in the simulation of section 4.2. Let the overall variation of Y_{ijr}^* based on model (5.12) be denoted by $Var\big(Y_{ijr}^*\big) = \sigma_y^2$. This variation has been set to $\sigma_y^2 = 0.36$, with time measured in years. Based on this and $\beta_0 = \log(24) = 3.178$, scenarios for simulation are considered with (i) proportions of explained variation of $EV = 1 - \sigma^2/\sigma_y^2 = 0.3$ and 0.5; (ii) correlation values between X_1 and X_2 of $\rho = 0$ and 0.3; and (iii) duration model coefficients $\beta_2 = 0$ and $\beta_1 = \beta_2$. The values of β_1 and β_2 can be obtained from EV and by noting that $\sigma_y^2 = Var(\alpha^*) + \beta_1^2 + \beta_2^2 + 2\rho\beta_1\beta_2 + \sigma^2$ when choosing, without loss of generality, $\mu_{x1} = \mu_{x2} = 0$ and $\sigma_{x1}^2 = \sigma_{x2}^2 = 1$.

From the eight possible scenarios given by (ii) and (iii) we explore four. These are presented in Table 5.1. The first two scenarios correspond to values of $EV = 0.5$ and $\sigma^2 = 0.18$ and the latter two, to $EV = 0.3$ and $\sigma^2 = 0.252$.

The "not jobless" spells were simulated according to an exponential distribution with mean $\gamma_1 \exp(\gamma_2 x_{i2})$, where $\gamma_1 = 11.619$ and $\gamma_2 = 0.155$; these values give a proportion of individuals with zero jobless spells in the population of about 59.5%.

Table 5.1: Parameter scenarios I-IV used for simulation for Kaplan-Meier estimation.

Scenario	I	II	III	IV
ρ	0.300	0.000	0.300	0.300
β_1	0.226	0.365	0.247	0.1535
β_2	0.226	0.000	0.000	0.1535
EV	0.5	0.5	0.3	0.3

The parameters of the loss to follow-up model in (5.13) were set up so that about 50% of the sample would drop out by year six, with $\alpha_0 = 2.131$ and $\alpha_1 = -0.536$. Since the dropout and the duration models share the variable X_2, $\alpha_1 < 0$ ensures that individuals in the sample with longer jobless spells are associated with a lower probability of being observed.

Four populations were generated according to the scenarios shown in Table 5.1, each of size $N = 100,000$. Within each population, there were 40.6% of individuals with at least one jobless spell (on average, over the four populations). Each population had individuals that experienced from zero to five spells.

Recall that 1000 samples were selected from the finite population, by selecting a random sample of size n_r from stratum r each time. Across scenarios I-IV and the 1000 samples, the average number of observed spells, the percentage of censored spells and of LTF individuals by year six did not vary substantially. On average, there were about 503 total jobless spells, 11.6% of those were censored, and 49.9% of persons were LTF. The average percentage (40.5%) of individuals with at least one jobless spell in the samples is representative of the average across populations.

The population quantity of interest is the distribution of the durations of the simulated jobless spells, expression (5.1). A summary of the survival distribution (duration times with probabilities closest to $.1, \ldots, .9$) from each population of the four is shown in Table A in the appendix.

Subsection 5.3.2 presents the formulas for variance estimation for the estimation methods: sampling design-weighted from the superpopulation and finite population

approaches (expressions (5.15), (5.16) and (5.17)), and the combined and combined naive (expressions (5.18) and (5.19)). These were carried out for each simulated scenario. Results for $\hat{S}(y)$ include bias, empirical and average standard errors, as well as estimated coverage probabilities of a 95% nominal confidence interval given by

$$\exp\Big\{-\exp\Big\{\hat{S}_l(y) \pm 1.96\sqrt{\hat{\mathrm{Var}}\big(\hat{S}_l(y)\big)}\Big\}\Big\}$$

where $\hat{S}_l(y) = \log\big(-\log\big(\hat{S}(y)\big)\big)$. This complementary log-log transformation of $\hat{S}(y)$ was used in order to have the CI limits between 0 and 1. General features of the results will be discussed below, and detailed results can be found in the Appendix (Tables A.2 - A.9).

Figures 5.1 and 5.2 show bias and estimated coverage from scenarios I-IV at each duration time where the estimated survival probability was closest to $0.1, \ldots, 0.9$. The graphs show results from the DES and COMB methods. Only the superpopulation design method is shown here, since the finite population methods gave very similar results (see appendix).

Results from scenario I ($\rho = 0.3, \beta_2 > 0$) show that in terms of bias and coverage, the COMB method does much better than DES, as expected. When $\rho = 0$ and $\beta_2 = 0$ as in scenario II, no IPC weights are needed, only design weights. This picture shows the COMB method does as well as DES, in terms of bias and coverage, even though IPC weights are not needed.

In scenario III, when $\rho = 0.3$ and $\beta_2 = 0$, we have that Y is independent of X_2 given X_1, thus IPC weights are not needed. Figure 5.2 shows that both DES and COMB give good results, with slightly bigger bias from DES.

The case of scenario IV is comparable to scenario I, only that the explained variation is decreased from 0.5 to 0.3. As with scenario I, the COMB method gives much better results in terms of both bias and coverage than the DES method.

In summary, through scenarios I-IV the COMB method gives the lowest bias and a coverage close to 0.95, and is never lower than 0.925. It has been observed that the average standard errors slightly underestimate the empirical standard errors from

the COMB method at some duration times, which gives a slightly lower coverage than the nominal 95%. The average standard errors estimate the empirical standard errors by ranges $91-96\%$, $92-98\%$, $93-102\%$, $93-101\%$ in scenarios I-IV, respectively (Tables A.2 to A.9). Another simulation using scenario I where the sample sizes were increased to 2900 (not shown here), gave slightly better results for the COMB method. The average standard errors estimate $93-100\%$ of the empirical standard errors and the coverage improves from ranges $0.925-0.947$ to $0.932-0.955$.

The COMB variance estimates give slightly lower values than the naive ones (Tables A.2-A.9), as expected (Robins et al. [51]), and CI coverage quite close to 95%. The use of the computationally simpler naive variance estimates provide conservative confidence inervals (coverage slightly greater than the nominal value), which are satisfactory in many practical settings. It is of interest to make a comparison between COMB naive and COMB estimates with real data, and this will be discussed in chapter 7, where methods are applied to SLID jobless spells.

In order to obtain unbiased results, COMB methods are advisable when it is suspected that the sampling design and dropout have an effect on durations. Even though the Naive COMB variance estimates are only slightly larger than the COMB variance estimates in this simulation, in practice, it is not recommended to rely only on the Naive COMB variance, but rather to compute both and see whether the COMB variance is substantially different. If this is the case, use of the COMB method is recommended.

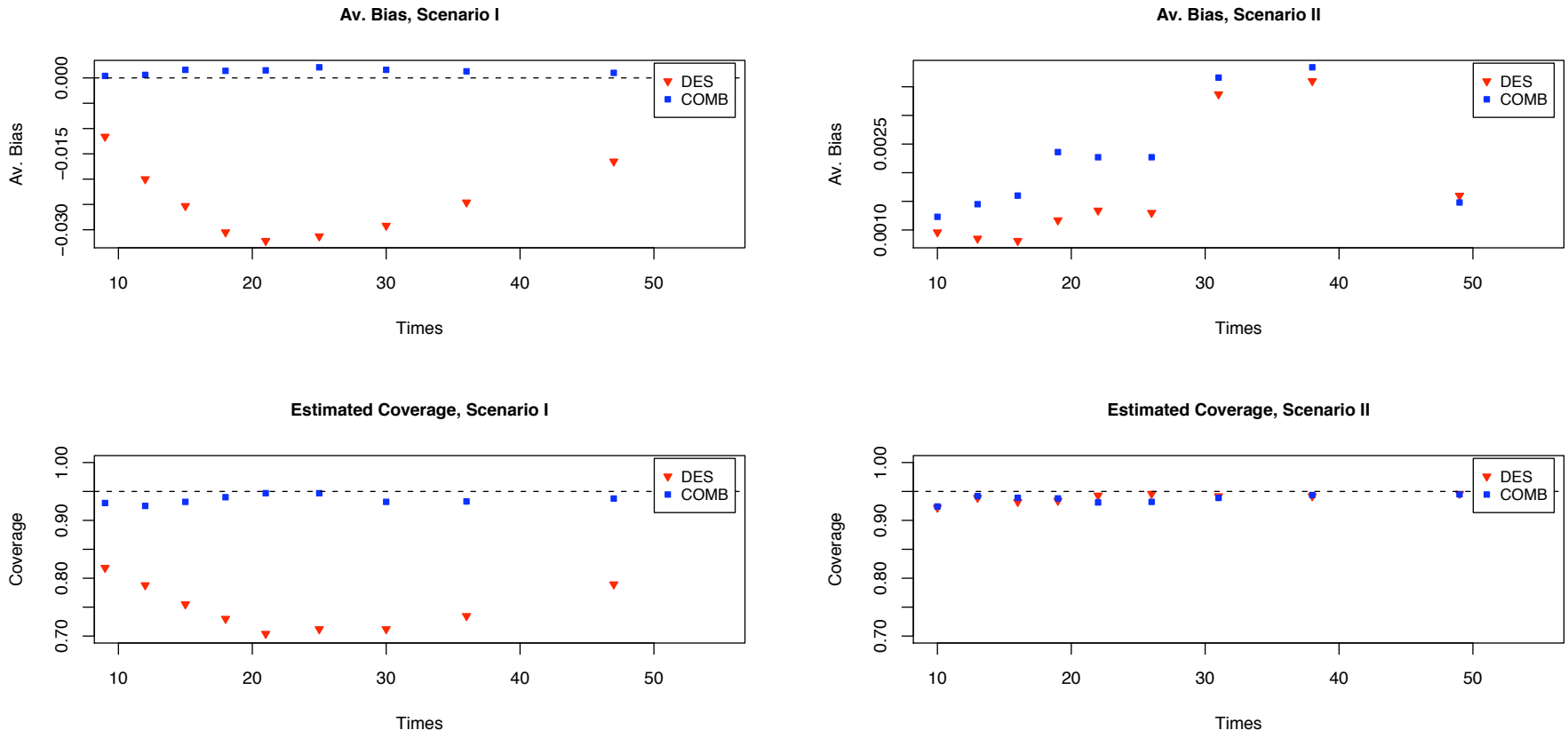

Figure 5.1: Bias and estimated coverage, scenarios I and II.

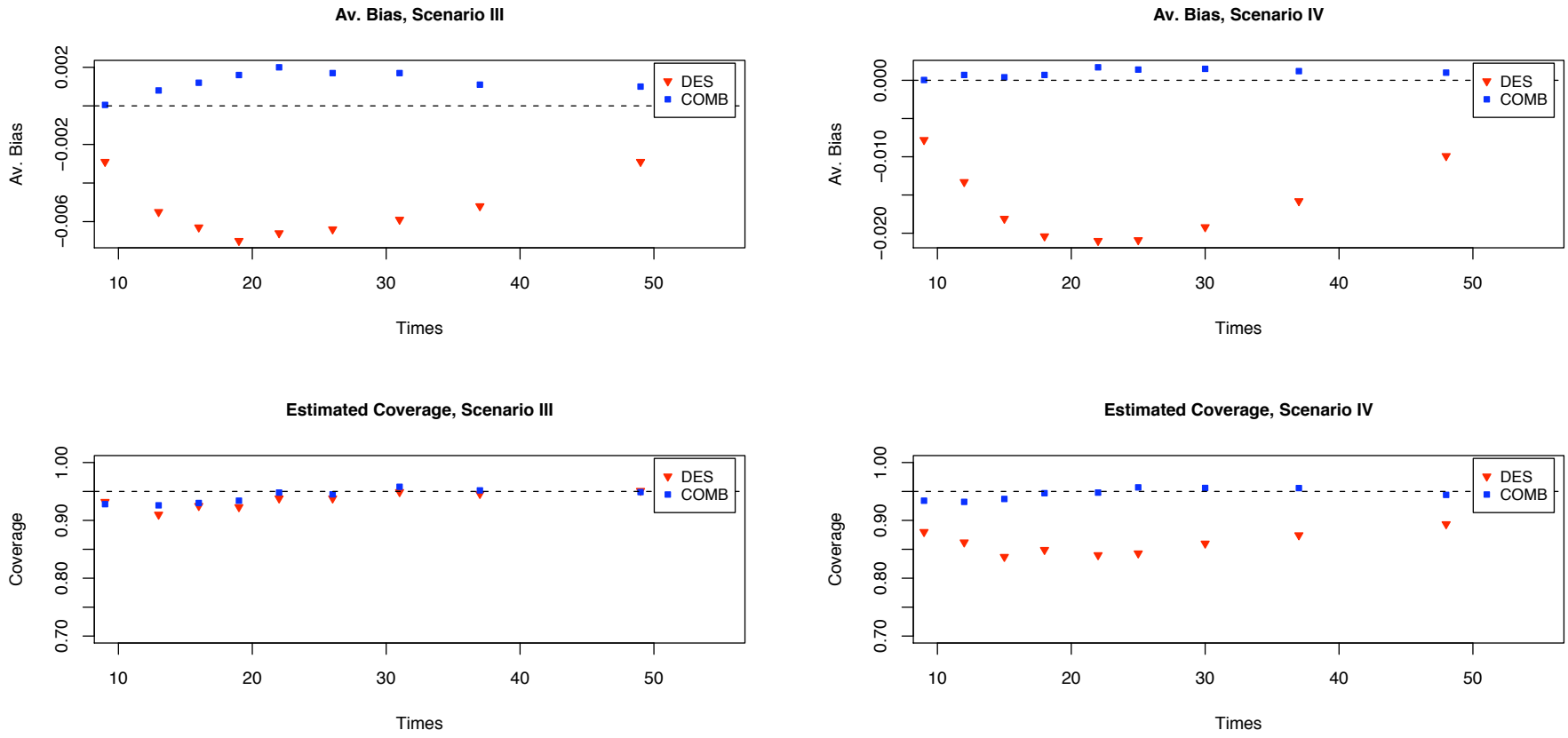

Figure 5.2: Bias and estimated coverage, scenarios III and IV.

Chapter 6

Weighted Cox PH Analysis

When analyzing spells and their relation with covariates, we distinguish cases where individuals experience a single spell, or cases where individuals experience a sequence of spells. Examples of single spells are the time to experience a developmental milestone in children, the time for women to have a first pregnancy, and their relationship with factors such as socio-economic status or family composition. Sequences of spells can be found for example in tobacco consumption studies, where sequences of tobacco smoking cessation periods with relation to education level and exposure to cigarette prevention campaigns are examined.

As mentioned before, methods for sampling design-weighted estimation of regression parameters based on the Cox model have been discussed by several authors, for example, Binder [4], Lin et al. [39], Boudreau and Lawless [9]. In the case of IPC weights, Robins and Finkelstein [50] used Cox PH model based methods in the analysis of data from a clinical trial to study the effect of an alternative treatment in AIDS patients. The estimating function formulas based on the Cox PH model presented in this chapter can be seen as a version of the formulas from Robins and Finkelstein developed for the context of duration analysis of survey data. Their variance estimation procedures, however, are complex and based on stochastic integrals and martingale theory; and do not consider sampling design weights, clustering or stratification.

As discussed earlier, the parametric methods for variance estimation presented in chapter 4 can be used when assuming that the duration time distribution is known except for a vector of parameters. They can also be used to estimate a discrete survivor distribution by viewing the Kaplan-Meier estimate as a maximum likelihood estimator, as discussed in chapter 5. The case of the Cox model is more complicated since the parameters are given by an infinite dimensional baseline hazard function $h_0(y)$ which is non-parametric, and a set of regression coefficients that are specified by a relative risk function, generally of the form $r(x, \beta) = \exp(\beta' x)$. One option is to approximate the Cox PH model with a piecewise constant (PC) hazards model. From this approach, the parametric based methods from chapter 4 can be employed.

The first section of this chapter gives expressions for the estimating functions that are used to estimate the regression coefficients and the cumulative hazard function based on the Cox PH model. It also provides a discussion on the piecewise constant model and its relationship with the Cox model. Section 6.2 gives a variance estimation method by adapting the variance estimation formulae from chapter 4 to the PC model. Finally, section 6.3 presents a short simulation in which this approximation is assessed.

6.1 IPC weights and the Cox PH model

6.1.1 Estimating functions

Recall from section 4.3 that when considering strata $r = 1, \ldots, R$ and clusters $k = 1, \ldots, K_r$ within stratum r, the estimating equations defined for individuals in the sample $S = \bigcup_{r=1}^{R} \bigcup_{k=1}^{K_r} S_{rk}$ are given by:

$$U(\theta, \alpha) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i \in S_{rk}} U_i(\theta, \alpha) = 0 \tag{6.1}$$

$$G(\alpha) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i \in S_{rk}} G_i(\alpha) = 0, \tag{6.2}$$

where the equations $U(\theta, \alpha)$ and $G(\alpha)$ involve the model for durations and dropout used to estimate the IPC weights, respectively.

Suppose that Y_i is a duration from individual i and that u_i and v_i denote the start and end times, respectively. Further, let $y_i = \min(Y_i, C_i - u_i)$ be the observed duration and $\delta_i = I(y_i = Y_i)$ indicate whether the duration is completely observed or censored. Furthermore, suppose that $Z_i(u_i + y)$ denotes a set of covariates for individual i, that may include information on prior event history up to time $t - 1$, where $t - 1 < u_i + y \leq t$, as well as on information on external covariates.

The Cox Proportional Hazards model is given by:

$$h(y|Z(u_i + y)) = h_0(y)\exp(Z(u_i + y)'\beta). \tag{6.3}$$

The estimating functions $U_i(\theta, \alpha)$ in (6.1) based on (6.3) can be seen as an extension of the formulas from Robins and Finkelstein [50] to our observation framework; or as an IPCW version of the expressions in Binder [4], Lin et al. [39], or Boudreau and Lawless [9]. They are given by:

$$U_i(\beta, \alpha) = w_i(y)\delta_i \left\{ Z_i(u_i + y_i) - \frac{S^{(1)}(y_i, \beta, \alpha)}{S^{(0)}(y_i, \beta, \alpha)} \right\}, \tag{6.4}$$

where

$$S^{(1)}(y, \beta, \alpha) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{j \in S_{kr}} w_j(y) I(y_j \geq y) Z_j(u_j + y) \exp(\beta' Z_j(u_j + y)), \tag{6.5}$$

$$S^{(0)}(y, \beta, \alpha) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{j \in S_{kr}} w_j(y) I(y_j \geq y) \exp(\beta' Z_j(u_j + y)), \tag{6.6}$$

where $w_i(y) = \sum_{t=1}^{M} R_{it}\xi_{it}(y)/\pi_i p_{it}(\alpha)$, and $\xi_{it}(y) = I(t - 1 < u_i + y \leq t)$. For convenience we write $w_i(y)$ to stand for $w_i(y; \alpha)$.

The estimating equations referring to the model for loss to follow-up have the same form as in (4.4). Estimates of (β, α) are obtained by solving these equations along with (6.1) with θ replaced with β and $U_i(\beta, \alpha)$ given by (6.4). The estimate of β can be readily obtained through Cox PH software that allows for case weights and left truncation, such as the `coxph` function in R/SPlus and the `PHREG` procedure

in SAS. Estimation is done by first estimating α and then using estimated weights $w_i(y, \hat{\alpha})$.

The input data frame usually needs to be arranged in the same way that was used for the example of alternating employment and unemployment processes in section 3.3. Table 3.1 shows the data frame in which information regarding to spells for employment and unemployment (E,U) is arranged. When analyzing one type of durations, we would focus on one of the lines corresponding to the process of interest (either E or U). In this Table, the variables "Start.w" and "Stop.w" are analogous to $y_i(t-1)$ and $y_i(t)$, and "Status" to $\delta_i(t)$. The covariates in the Table shown as"z_{i1}", "z_{i2}", etc., are equivalent to $Z_i(1)$, $Z_i(2)$, etc. Such software also produces variance estimates assuming the weights as fixed. In many cases, the variance estimates assuming fixed weights are just slightly larger than those which recognize the IPC weights are random (Robins et al. [51]). This has been seen in the simulation results from chapters 4 and 5 regarding parametric models and Kaplan-Meier estimation, respectively, and will be assessed in the simulation study for the Cox PH model in section 6.3.

The estimate of the cumulative hazard function is given by

$$\hat{\Lambda}_0(y) = \sum_{y_i \le y} \left\{ \frac{\delta_i \hat{w}_i(y_i)}{\sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{j \in S_{rk}}^{n} \hat{w}_j(y_i) I(y_j \ge y_i) \exp(\hat{\beta}' Z_j(u_j + y_i))} \right\}. \quad (6.7)$$

Stratified versions of (6.3) and (6.7) can also be considered. For example, SLID has multiple levels of stratification within provinces, such as economic regions, employment insurance regions, etc. While modelling jobless spells from Ontario, one might be interested in including some of these as strata in the hazards model in (6.3) and so the model may, for instance, be expressed as

$$h_r(y|Z_r(t)) = h_{0r}(y)\exp(Z_r(t)'\beta). \quad (6.8)$$

Model strata is readily implemented in Cox model software, for example, by adding a term `strata(variable)` in the `coxph` function in R/SPlus or by defining a `strata` step inside the SAS procedure `PROC PHREG`.

6.1.2 Piecewise constant approximation

The piecewise constant (PC) hazards model can be used to approximate the Cox PH model. This model uses the form in (6.3) but with a piecewise constant specification for $h_0(y)$. It approximates results from a Cox model quite well, with a few well chosen pieces, and variance estimates based on the methods described in chapter 4 can be used. Information about the PC model can be found in Lawless [32] (pages 30, 323, 384) and some examples of applications of this model as an approximation to the Cox PH model in Lawless et al. [35], and Andersen et al.[2].

Let us follow the notation used in section 4.1, where u_i and v_i are the start and end times of a spell, and the spell's duration is given by $Y_i = v_i - u_i$, for individuals $i = 1, ..., n$. Let $b_0 < b_1 \cdots < b_m$ be specified values with $b_0 = 0$ and $b_m = \infty$. The hazard function for the PC model, given covariates, is then

$$h^{pc}(y|Z_i(u_i + y)) = h_0^{pc}(y) \exp(\beta' Z_i(u_i + y)), \tag{6.9}$$

where the piecewise constant baseline hazard function $h_0^{pc}(y)$ for y is:

$$h_0^{pc}(y) = \rho_j, \text{ if } \; b_{j-1} \leq y < b_j, \tag{6.10}$$

where $\rho_j > 0$ and $j = 1, 2, ..., m$.

The hazard function in (6.10) can be used to express the log-likelihood function for right censored data accounting for the time-varying IPC weights (see expression (4.3)). For simplicity, let's assume that the covariates are constant over intervals $(t-1, t]$, so that $Z(u_i + y) = Z(t)$. Let $y_i(t) = \min(t, v_i) - \min(t, u_i)$ be the length of the observed duration y_i at t and $\delta_i(t) = I(t - 1 < v_i \leq t)$ indicate whether the duration y_i ends in the interval $(t-1, t]$. Further, let $\theta = (\rho', \beta')'$, denote the parameter of interest, where $\rho = (\rho_1, \ldots, \rho_k)'$, $\beta = (\beta_1, \ldots, \beta_p)'$. The log-likelihood is of the following form:

$$l(\theta, \alpha) = \sum_{i=1}^{n} \sum_{t=1}^{M} w_{it}(\alpha) \; l_{it}(\theta), \tag{6.11}$$

where $w_{it}(\alpha) = R_{it}/\pi_i \; p_{it}(\alpha)$ and

$$\begin{aligned} l_{it}(\theta) &= \delta_i(t) \log h\big(y_i(t)\big) + \log\Big(S\big(y_i(t)\big)/S\big(y_i(t-1)\big)\Big) \\ &= \delta_i(t)\Big[\sum_{j=1}^{m} I(b_{j-1} \le y_i(t) < b_j) \log \rho_j + \beta' Z_i(t)\Big] \\ &\qquad - e^{\beta' Z_i(t)}\Big[\Lambda_0^{pc}(y_i(t)) - \Lambda_0^{pc}(y_i(t-1))\Big], \end{aligned} \tag{6.12}$$

where $\Lambda_0^{pc}(y) = \sum_{j=1}^{m} \rho_j \Delta_j(y)$ and $\Delta_j(y) = \int_{b_{j-1}}^{b_j} I(u \le y) du = \min(b_j, y) - \min(b_{j-1}, y)$.

Putting aside strata and clusters for now, the estimate $\hat{\theta} = (\hat{\rho}', \hat{\beta}')'$ is obtained from solving the system

$$U(\theta, \alpha) = \sum_{i=1}^{n} \sum_{t=1}^{M} U_{it}(\theta, \alpha) = 0 \tag{6.13}$$

$$G(\alpha) = \sum_{i=1}^{n} \sum_{t=1}^{M} G_{it}(\alpha) = 0 \tag{6.14}$$

where

$$U_{it}(\theta, \alpha) = w_{it}(\alpha) \frac{\partial l_{it}(\theta)}{\partial \theta}, \tag{6.15}$$

$w_{it}(\alpha) = R_{it}/\pi_i p_{it}(\alpha)$, and from (6.12),

$$\frac{\partial l_{it}(\theta)}{\partial \beta} = \delta_i(t) Z_i(t) - Z_i(t) \exp(\beta' Z_i(t)) \sum_{j=1}^{m} \rho_j [\Delta_j(y_i(t)) - \Delta_j(y_i(t-1))] \tag{6.16}$$

$$\frac{\partial l_{it}(\theta)}{\partial \rho_j} = \delta_i(t) \frac{I(b_{j-1} \le y_i(t) < b_j)}{\rho_j} - \exp(\beta' Z_i(t))[\Delta_j(y_i(t)) - \Delta_j(y_i(t-1))], \tag{6.17}$$

and where $G_{it}(\alpha)$ is the score estimating function from the model for $\text{logit}\big(\lambda_{it}(\alpha)\big) = \text{logit}\big\{Pr\big(R_{it} = 1 | R_{i,t-1} = 1, Z_i^c(t)\big)\big\}$ in chapter 4, expression (4.4).

The estimate $\hat{\theta} = (\hat{\rho}', \hat{\beta}')'$ can be obtained by maximizing $\sum_{i=1}^{n} \sum_{t=1}^{M} w_{it}(\hat{\alpha}) \; l_{it}(\rho, \beta)$ or $\sum_{i=1}^{n} \sum_{t=1}^{M} w_{it}(\hat{\alpha}) \; l_{it}(\tilde{\rho}(\beta, \hat{\alpha}), \beta)$ in (6.12) using optimization software (for example, `nlm` or `nlmin` in R/SPlus). Here, $\tilde{\rho}(\beta, \hat{\alpha})$ is the maximizer of (6.13) with α replaced by $\hat{\alpha}$ and β fixed and is given by

$$\tilde{\rho}_j(\beta, \hat{\alpha}) = \frac{\sum_{i=1}^{n} \sum_{t=1}^{M} w_{it}(\hat{\alpha}) \delta_i(t) I(b_{j-1} \le y_i(t) < b_j)}{\sum_{i=1}^{n} \sum_{t=1}^{M} w_{it}(\hat{\alpha}) \exp(\beta' Z_i(t))[\Delta_j(y_i(t)) - \Delta_j(y_i(t-1))]}. \tag{6.18}$$

Estimates for (6.18) are readily obtained by substituting β by $\hat{\beta}$. The baseline cumulative hazard function has the form

$$\hat{\Lambda}_0^{pc}(y) = \sum_{j=1}^{m} \hat{\rho}_j \Delta_j(y), \tag{6.19}$$

where $\hat{\rho}_j = \tilde{\rho}_j(\hat{\beta}, \hat{\alpha})$.

We can obtain an expression for the profile score $U^P(\beta, \alpha)$ for β based on (6.13) by substituting $\tilde{\rho_j}(\beta, \alpha)$ in place of ρ_j in expression (6.11), and differentiating with respect to β. The expression for $U^P(\beta, \alpha)$ has a similar form to the Cox model estimating function in (6.4) and its components in (6.5) and (6.6). It is given by

$$U^P(\beta, \alpha) = \sum_{i=1}^{n} U_i^P(\beta, \alpha), \tag{6.20}$$

where

$$U_i^P(\beta, \alpha) = \sum_{t=1}^{M} w_{it}(\alpha)\delta_i(t) \left\{ Z_i(t) - \sum_{j=1}^{k} I(b_{j-1} \leq y_i(t) < b_j) \frac{S_j^{(1)}}{S_j^{(0)}} \right\}, \tag{6.21}$$

$$S_j^{(0)} = \sum_{i=1}^{n} \sum_{t=1}^{M} w_{it}(\alpha) \exp(\beta' Z_i(t))[\Delta_j(y_i(t)) - \Delta_j(y_i(t-1))], \tag{6.22}$$

$$S_j^{(1)} = \sum_{i=1}^{n} \sum_{t=1}^{M} w_{it}(\alpha) Z_i(t) \exp(\beta' Z_i(t))[\Delta_j(y_i(t)) - \Delta_j(y_i(t-1))], \tag{6.23}$$

where as before, $w_{it}(\alpha) = R_{it}(y)/\pi_i p_{it}(\alpha)$.

Introducing the sampling design's stratum and cluster information we have an analogous version of (6.18) given by

$$\tilde{\rho}_j^s(\beta, \hat{\alpha}) = \frac{\sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i \in S_{rk}} \sum_{t=1}^{M} w_{it}(\hat{\alpha}) \delta_i(t) I(b_{j-1} \leq y_i(t) < b_j)}{\sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i \in S_{rk}} \sum_{t=1}^{M} w_{it}(\hat{\alpha}) \exp(\beta' Z_i(t))[\Delta_j(y_i(t)) - \Delta_j(y_i(t-1))]}. \tag{6.24}$$

The profile score function that accounts for clusters and strata can be obtained by replacing $\sum_{i=1}^{n} U_i^P(\beta, \alpha)$ by $\sum_{r=1}^{R} \sum_{k=1}^{Kr} U_i^P(\beta, \alpha)$ in expression (6.20), and by doing the same with the summations in its components $S_j^{(0)}$ and $S_j^{(1)}$ in (6.22) and (6.23).

When k becomes large, in practice, we'll assume that b_{k-1} is fixed at some large value beyond which a failure is impossible; and that as k increases, the values $b_j - b_{j-1}$ for $j = 1, \ldots k-1$ approach 0. Then the expression for $U_i^P(\beta, \alpha)$ based on the PC model in (6.21) converges to $U_i(\theta, \alpha)$ based on the Cox model, in (6.4). Note that their respective components $S_j^{(0)}$ and $S_j^{(1)}$ in (6.22) and (6.23) expressed accounting for clusters and strata converge to $S^{(0)}(y, \beta, \alpha)$ and $S^{(1)}(y, \beta, \alpha)$ in (6.5) and (6.6), respectively (see Lawless [32], p. 385). Analogously, the baseline cumulative hazard function in (6.7) is approximated by

$$\hat{\Lambda}_0^{pc}(y) = \sum_{j=1}^{m} \hat{\rho}^s \Delta_j(y). \tag{6.25}$$

where $\hat{\rho}^s = \tilde{\rho}_j^s(\hat{\beta}, \hat{\alpha})$.

6.2 Variance estimation

The variance estimates based on the PC model can be obtained simply by applying the formulae described in Chapter 4, in expression (4.26) given by

$$\widehat{Var}(\hat{\theta})_{comb} = \hat{B}_{comb}^{-1} \hat{C}_{comb} \hat{B}_{comb}^{-1\,\prime}, \tag{6.26}$$

where

$$\hat{B}_{comb} = -\sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i \in S_{rk}} \frac{\partial U_i(\theta, \hat{\alpha})}{\partial \theta}\Big|_{\hat{\theta}} \quad \text{and} \tag{6.27}$$

$$\hat{C}_{comb} = \sum_{r=1}^{R} \frac{K_r}{K_r - 1} \sum_{k=1}^{K_r} \left\{ \left(\sum_{i \in S_{rk}} \hat{E}_i \right) - \left(\frac{1}{K_r} \sum_{k=1}^{K_r} \sum_{i \in S_{rk}} \hat{E}_i \right) \right\}^{\otimes 2} \tag{6.28}$$

where $A^{\otimes 2} = AA'$, and

$$\hat{E}_i = U_i(\hat{\theta}, \hat{\alpha}) - \left[\sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i \in S_{rk}} U_i(\hat{\theta}, \hat{\alpha}) G_i(\hat{\alpha})' \right] \left[\sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i \in S_{rk}} G_i(\hat{\alpha}) G_i(\hat{\alpha})' \right]^{-1} G_i(\hat{\alpha}).$$

where $\theta = (\beta', \rho')'$ is the parameter vector of the PC model in (6.11), and is of dimension $(p + m)$. Here, $U_i(\theta, \alpha) = \sum_{t=1}^{M} U_{it}(\theta, \alpha)$, where the summands are based on the PC model and are given in (6.15). As mentioned earlier regarding the estimating equation in (6.14), the vector $G_i(\alpha) = \sum_{t=1}^{M} G_{it}(\alpha)$, is of dimension $q = q_1 + \cdots + q_M$.

The $(p + m) \times (p + m)$ matrix $\hat{B}$ has the form

$$\hat{B} = -\begin{pmatrix} \partial U(\theta, \alpha)/\partial\beta\partial\beta' & \partial U(\theta, \alpha)/\partial\rho'\partial\beta' \\ \partial U(\theta, \alpha)/\partial\beta\partial\rho' & \partial U(\theta, \alpha)/\partial\rho\partial\rho' \end{pmatrix}\Big|_{\hat{\theta}}, \tag{6.29}$$

where, by letting $Z_{ik}(t)$ and β_k be the kth column of $Z_i(t)$ and the kth component of β, from (6.16) and (6.17) we have

- $\partial U(\theta, \alpha)/\partial\beta\partial\beta'$ is a $p \times p$ matrix with (k, s) elements of the form

$$\partial U(\theta, \alpha)/\partial\beta_k\partial\beta_s = -\sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i \in S_{rk}} \sum_{t=1}^{M} \exp\left(\beta' Z_i(t)\right) Z_{ik}(t)' Z_{is} \Big[\Lambda_0^{pc}\big(y_i(t)\big) - \Lambda_0^{pc}\big(y_i(t-1)\big) \Big]$$

for $k = 1, \ldots, p$, $s = 1, \ldots, p$;

- $\partial U(\theta, \alpha)/\partial\beta'\partial\rho'$ is a $p \times m$ matrix with (k, j) elements

$$\partial U(\theta,\alpha)/\partial\beta_k\partial\rho_j = -\sum_{r=1}^{R}\sum_{k=1}^{K_r}\sum_{i\in S_{rk}}\sum_{t=1}^{M}\exp\big(\beta' Z_i(t)\big)Z_{ik}(t)\Big[\Delta_j\big(y_i(t)\big) - \Delta_j\big(y_i(t-1)\big)\Big]$$
$$= \partial U(\theta,\alpha)/\partial\rho_j\partial\beta_k, \text{ the } (j,k) \text{ element of } \partial U(\theta,\alpha)/\partial\rho'\partial\beta',$$

where $k = 1, \ldots, p$ and $j = 1, \ldots, m$;

- $\partial U(\theta, \alpha)/\partial\rho'\partial\rho'$ is a $m \times m$ diagonal matrix with (j, j') elements

$$\partial U(\theta,\alpha)/\partial\rho_j\partial\rho_{j'} = -\sum_{r=1}^{R}\sum_{k=1}^{K_r}\sum_{i\in S_{rk}}\sum_{t=1}^{M} w_{it}\ \Big[\delta_i(t)I(b_{j-1} \le y_i(t) < b_j)/\rho_j^2\Big]$$

if $j = j'$ and equal to zero if $j \neq j'$, $j = 1, \ldots, m$.

6.3 A simulation study

6.3.1 Setup

The simulation presented here aims to assess the PC approximation to the Cox PH model. We assume an observational framework where spells from individuals in a finite population of size $N = 100,000$ are simulated. Simple random samples of $n = 1500$ individuals are obtained from this population and follow-up is simulated annually for six consecutive years. The observation period is denoted by $(0, M]$ where $M = 6$ and interview times are given by t, $t \in \{1, 2, \ldots, 6\}$. Each individual i in the population is associated with one simulated duration Y_i, where Y_i has a Weibull distribution, that is,

$$\big(Y_i^*|x_{i1}, x_{i2}\big) = \big(\log Y_i|x_{i1}, x_{i2}\big) \sim \text{Extreme Value } \big(u(x_i), b\big) \tag{6.30}$$

where $u(x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$; and $(X_1, X_2) \sim$ Bivariate Normal with vector mean $\mu = (\mu_{x_1}, \mu_{x_2})$ and a variance matrix with elements $Var(X_1) = \sigma^2_{x_1}$, $Var(X_2) = \sigma^2_{x_2}$ and covariance $\text{Cov}(X_1, X_2) = \sigma_{x_1,x_2}$. The starting times of the durations were generated from a Uniform distribution in $(0, 3)$ years.

Simulating the log-spell durations $(Y_i^*|x_{i1}, x_{i2})$ from an Extreme Value distribution implies that the variables $(Y_i|x_{i1}, x_{i2})$ have a Weibull distribution with shape and scale parameters $\exp\big(u(x_i)\big)$ and b^{-1}, respectively. The Weibull distribution has the peculiarity that it belongs to both the Accelerated Failure Time and to the Proportional Hazards families of distributions (see more on Weibull and Extreme Value distributions in Lawless [32], sections 5.2 and 6.3).

The start times of the simulated spells follow a Uniform$(0, 3)$ distribution. As in previous simulations, loss to follow-up is simulated from a logistic model, where

$$\text{logit}\lambda_{it}(\alpha) = \alpha_0 + \alpha_1 x_{i2} \quad t = 1, \ldots, 5. \tag{6.31}$$

with $\lambda_{it}(\alpha) = \Pr\left(R_{it}|R_{i,t-1} = 1, X_{i2} = x_{i2}\right)$ and from where the IPC related probabilities are obtained: $p_{it}(\alpha) = \Pr\left(R_{it=1}|x_{i2}\right) = \lambda_{i1}\lambda_{i2}\cdots\lambda_{it}$. The estimation of the IPC weights was done by fitting a separate model for each time $t \in \{1, \ldots, 5\}$, where $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_5)'$, where $\hat{\alpha}_t = (\hat{\alpha}_{0t}, \hat{\alpha}_{1t})$.

The Cox PH and PC models we consider are given, respectively, by

$$h^c(y|x_1) = h_0^c(y)\exp(\beta_1^c x_1) \tag{6.32}$$

and

$$h^{pc}(y|x_1) = h_0^{pc}(y)\exp(\beta_1^{pc} x_1). \tag{6.33}$$

That is, x_1 is included in the fitted models. The "targets" of estimation are then $\beta_{1_N}^c$ and $\beta_{1_N}^{pc}$, which are solutions to the following estimating equations from the entire simulated population. The estimating equations based on the Cox model are given by

$$U(\beta_1^c) = \sum_{i=1}^{N} \delta_i \Big\{ x_{i1} - \frac{S^{(1)}(y_i, \beta_1^c)}{S^{(0)}(y_i, \beta_1^c)} \Big\} \tag{6.34}$$

where

$$S^{(1)}(y, \beta_1^c) = \sum_{j=1}^{N} I(y_j \geq y) x_{j1} \exp(\beta_1^c x_{j1}), \tag{6.35}$$

$$S^{(0)}(y, \beta_1^c) = \sum_{j=1}^{N} I(y_j \geq y) \exp(\beta_1^c x_{j1}); \tag{6.36}$$

where $y_i = \min(Y_i, M - u_i)$ is the spell's length in $(0, M)$ and $\delta_i = I(y_i \leq M - u_i)$. The estimating equations based on the PC model are given by

$$U(\beta^{pc}) = \sum_{i=1}^{N} \Big\{ \delta_i x_{i1} - x_{i1} \exp\big(\beta_1^{pc} x_{i1}\big) \sum_{j=1}^{m} \rho_j \Delta_j\big(y_i\big) \Big\}. \tag{6.37}$$

Note that (6.36) and (6.37) do not need to account for weights (design or IPC) since they are from the entire population.

The parameter values for the model in (6.30) were selected according to a total variation of $\sigma_y^2 = Var(Y_i^*) = 0.36$. The model in (6.30) can be expressed as

$$Y^* = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \text{ where } \epsilon \sim \text{Extreme Value}(0, b)$$

and the total variation is expressed as $\sigma_y^2 = \beta_1^2 + \beta_2^2 + 2\beta_1\beta_2 Corr(X_1, X_2) + \sigma_\epsilon^2$, where $\sigma_\epsilon^2 = Var(\epsilon) = b^2\pi^2/6$, which is unexplained variation.

The value of $\beta_0 = 3.369$ was selected from the relation $\beta_0 = \beta_0' + \kappa\sigma'\sqrt{6}/\pi$, where $\kappa = 0.5772$ is known as Euler's constant (see Lawless [32], p.21), and where $\beta_0' = \log(24) = 3.178$ (with durations measured in weeks) and $\sigma' = \sqrt{0.18} = 0.424$ are the intercept and standard deviation that were used to simulate log-Normal durations in the simulation from section 5.3. This value for β_0 gives a similar duration distribution to the one used before, with a median duration value of 24 weeks.

The proportion of unexplained variation was set to $EV = 0.5$, so $Var(\epsilon) = .18$. Without loss of generality, $\mu_{x1} = \mu_{x2} = 0$ and $\sigma_{x1}^2 = \sigma_{x2}^2 = 1$, and $Corr(X_1, X_2) = 0.3$. Based on this and $\beta_0 = 3.369$, the remaining parameter values for the duration model were obtained as $\beta_1 = \beta_2 = 0.263$ and $b = 0.331$. The loss to follow-up model parameters in (6.31) were set as $\alpha_0 = 2.131$ and $\alpha_1 = -0.537$, so that 50% of the of the samples would drop out by year six.

6.3.2 Results

A total of 1000 simple random samples of size $n = 1500$ were drawn from the population and estimates of β_{1N}^c and β_{1N}^{pc} were obtained and denoted by $\hat{\beta}_{h1}^c$ and $\hat{\beta}_{h1}^{pc}$, for $h = 1, \ldots, 1000$. The objective is to compare $\bar{\hat{\beta}}_1^c = \sum_{h=1}^{1000} \hat{\beta}_{h1}^c/1000$ and $\bar{\hat{\beta}}_1^{pc} = \sum_{j=1}^{1000} \hat{\beta}_{j1}^{pc}/1000$ with β_{1N}^c and β_{1N}^{pc} and examine their variance estimates.

The PC model intervals $(b_{j-1}, b_j]$ were selected using the same procedure for the population and samples, based on the durations distribution, as follows. The middle point between the minimum duration and the first quartile was found, then between the first and the second, and between the second quartile and the third. These amount for 6

pieces so far and then there were included 6 additional pieces of the same length between the maximum duration and the third quartile, giving a total of 12 pieces. This method is artificial, but was found convenient for simulation purposes and was the option that gave the best results. With real data, it is advisable rather to determine the length and number of pieces based on the curvature of the baseline cumulative hazard function from the Cox model. In a way, the cumulative baseline function serves as a tool for "calibration" of the PC model in order to specify the pieces that yield a better approximation.

The estimation procedure for the pieces ρ_j for $j = 1, \ldots, m$ and the parameters β_{1N}^{pc} was performed as described in section 6.1.2, using the R non-linear minimization function `nlm`. The graph in the left hand side in Figure 6.1 shows the baseline cumulative hazard function based on the durations from the entire population, obtained from the Cox PH and the PC models, respectively. The right hand side graph shows the average of the unweighted estimates of the baseline cumulative hazard functions from the 1000 samples, restricted to up to 80 weeks, based on the Cox PH and the PC models. The IPC estimates gave very similar results and were omitted from the graph.

It can be seen that the Cox PH model is approximated fairly closely by the PC model in the population case. In the sample based average estimate, the PC model starts disagreeing from the Cox model for durations of around 52 weeks, which should not be of concern, since durations longer than 52 weeks have a low probability of occurring, that is, $Pr(Y_i > 52) = 0.082$ (calculated from the population duration distribution). Based on these graphs, we can conclude that the PC approximation in both the population and the samples is quite good. This is further verified with the point and variance estimation results presented in the following discussion.

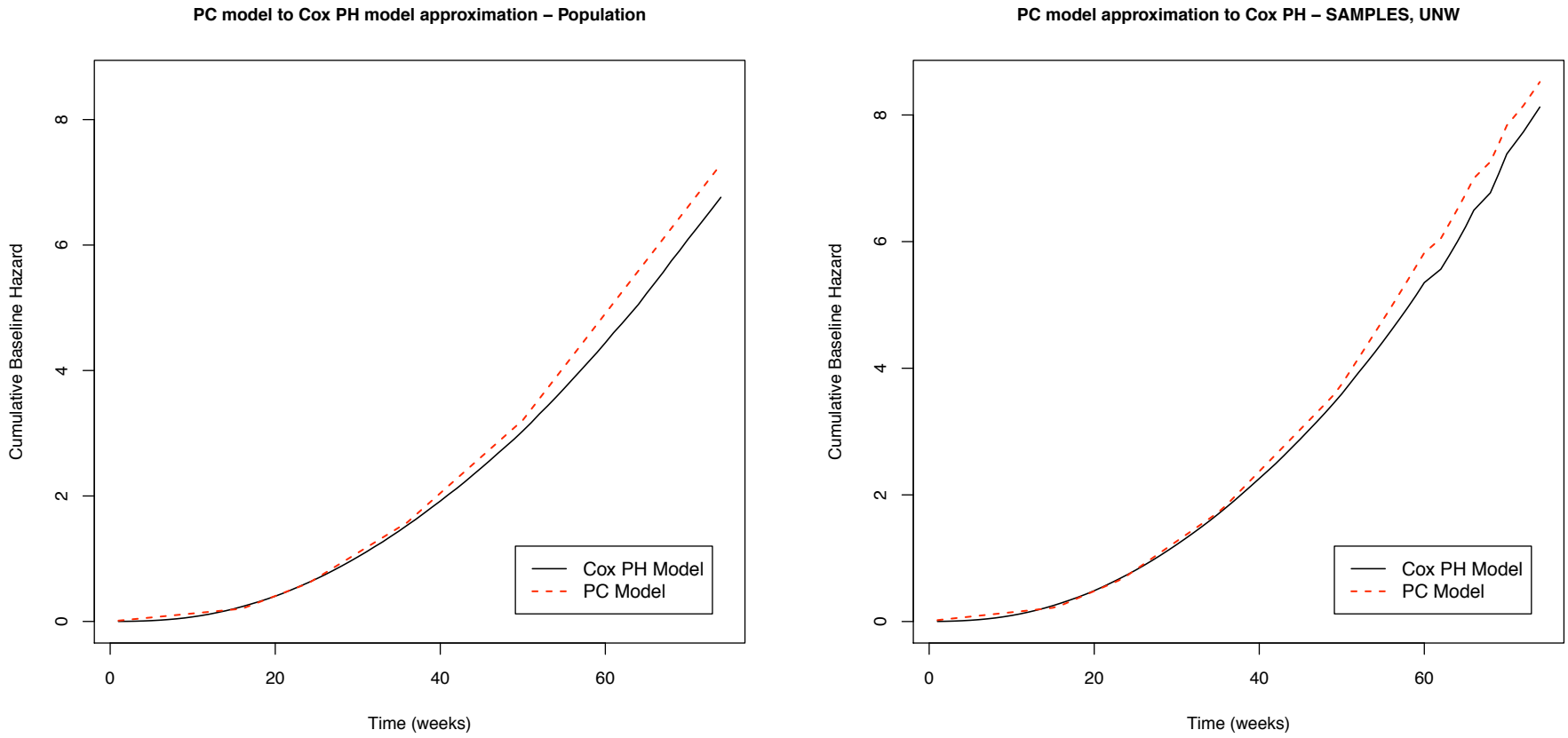

Figure 6.1: Population and sample average of the baseline cumulative hazard function.

In all samples, some individuals were loss to follow-up at year 1 and their spell was not observed. On average, there were 49.9% individuals lost to follow-up and a total of 1182 observed spells, from which 6.5% were censored.

Under the full model $\left(Y^*|X_1, X_2\right)$, the relationship between the regression coefficients from the Extreme Value model β_1 in (6.30) and the Cox PH model β_1^{ph} is given by $\beta_1^{ph} = -\beta_1/b$, where b is the scale parameter from the Extreme Value model. Note that if we were using the full model, we would have $\beta_1^{ph} = -0.795$, since we used $b = 0.331$ and $\beta_1 = 0.263$. The value for β_1^c under the reduced model $[Y^*|X2]$ should be similar to β_1^{ph}. Table 6.2-(a) shows the values of the target population parameters based on the Cox PH and PC models denoted earlier by $\beta_{1_N}^c$ and $\beta_{1_N}^{pc}$, respectively; the latter resulting a good approximation of the former. This section of the table also shows the average over the 1000 sample estimates, based on the Cox PH and PC model, which are also very close to one another.

Table 6.1 shows the estimated coverage under the Cox PH model with respect to the target parameter from this same model is 0.955 and 0.941, respectively, from the unweighted and IPC methods. Regarding the PC model coverage, it was estimated as 0.961 and 0.959 from the unweighted and IPC methods, respectively. We note that in this particular example, there was not a strong effect of loss to follow-up.

Table 6.1: Estimated coverage based on estimates from Cox PH and PC models.

Model	Unw	IPC
Cox PH	0.955	0.941
PC	0.961	0.959

Results shown in Table 6.2-(b) give an indication of how well the variance estimation based on the PC model approximates the variance based on the Cox PH model. The second column, labeled as "Cox PH", shows the average standard errors over the 1000 samples, based on the conservative Cox PH model variance, given by $\mathcal{I}(\hat{\beta}_1^c)^{-1}$, where

$$\mathcal{I}(\hat{\beta}_1^c) = -\frac{\partial U(\beta_1^c)}{\partial \beta_1^c}\Big|_{\hat{\beta}_1^c} = -\sum_{i=1}^{n} \frac{\partial U_i(\beta_1^c)}{\partial \beta_1^c}\Big|_{\hat{\beta}_1^c} \tag{6.38}$$

$$U_i(\beta_1^c) = w_i(y)\delta_i \left\{ x_{i1} - \frac{S^{(1)}(y_i, \beta_1^c)}{S^{(0)}(y_i, \beta_1^c)} \right\}, \tag{6.39}$$

Table 6.2: Population target parameter and estimates based on the Cox PH and PC models.Samples are of size 1500 and subject to loss to follow-up.

(a) Point estimates

	Population	Sample Average	
	$(\beta^{c}_{1_N},\ \beta^{pc}_{1_N})$	Unw	IPC
Cox PH	-0.7745	-0.7816	-0.7807
PC	-0.7736	-0.7828	-0.7767

(b) Standard errors of the form $1/I(\hat{\theta})$

Weights	Cox PH	PC,nlm	PC,coded
Unity	0.0361	0.0353	0.0353
IPC	0.0312	0.0310	0.0310

(c) Standard errors from sandwich variance estimates

	Unweighted		IPC		IPC Naive*
Model	Av.se	Emp.se	Av.se	Emp.se	Av.se
Cox PH	0.0382	0.0371	NA	0.0434	0.0425
PC	0.0380	0.0371	0.0420	0.0444	0.0426

* Treats IPC weights as fixed.

and

$$S^{(1)}(y,\beta_1^c) = \sum_{i=1}^{n} w_i(y) I(y_j \geq y) x_{i1} \exp(\beta_1^c x_{i1}),$$
$$S^{(0)}(y,\beta_1^c) = \sum_{i=1}^{n} w_i(y) I(y_j \geq y) \exp(\beta_1^c x_{i1}), \tag{6.40}$$

and where $w_i(y) = 1$ and $w_i(y) = \sum_{t=1}^{M} R_{it} I(t-1 \leq u_i + y < t)/p_{it}(\hat{\alpha})$ were used in the unity and IPC weighted case, respectively, where the latter are considered as fixed.

The third and fourth columns in Table 6.2-(b), labeled as "PC,nlm" and "PC,coded" correspond to the PC model based standard errors. The former is computed as the inverse of the negative hessian matrix given by the optimization R function `nlm`; the latter is based on the analytically coded hessian matrix, where the second derivatives were coded according to (6.29). In both cases, unweighted and IPC, the approximations are very good. The "PC,nlm" and the "PC,coded" standard errors are virtually identical and they closely approximate the standard errors based on the Cox PH model, for both unity and IPC weights, with an average about 2% of the Cox PH average. It should be noted that these standard errors are not correct, but are shown to indicate how well the PC approximation works regarding $\mathcal{I}(\hat{\beta})^{-1}$. For unweighted estimates, the sandwich and $\mathcal{I}(\hat{\beta})^{-1}$ variance estimates should be close.

Finally, Table 6.2-(c) shows the average and empirical standard errors over the 1000 samples, based on Cox PH and PC sandwich variance estimates for unity and IPC weights. The second and third columns labeled as "Av.se" and "Emp.se" correspond to those where unity weights were used. The variance based on the Cox PH model was calculated from the sandwich form

$$\mathcal{I}(\hat{\beta}_1^c)^{-1}\Big(\sum_{i=1}^{n} U_i^s(\hat{\beta}_1^c) U_i^s(\hat{\beta}_1^c)'\Big)\mathcal{I}(\hat{\beta}_1^c)^{-1} \tag{6.41}$$

where $\mathcal{I}(\hat{\beta}_1^c)$ as in (6.38) and $U_i^s(\beta_1^c)$ are score residuals from the Cox PH model, obtained using the `residual` function together with the `scores` option in R. This variance formula with unity weights gave an average standard deviation of 0.0382 which overestimates slightly, of about 3% the empirical value, 0.0371. The unweighted variance based on the PC model was obtained as described in section 6.2, simplified by the absence of stratification and clustering. The PC model standard errors are given by 0.0380 and 0.0371, and indicate that the PC model gives a very good approximation to the sandwich

variance estimates from the Cox PH model, when unity weights are used.

The IPCW case is shown in fourth, fifth and sixth columns in Table 6.2-(c). The first two were calculated based on (6.41) and the variance formulas from section 6.2, respectively. The average standard error from the PC model is equal to 0.0420, which estimates 94.5% of the empirical value 0.0444 and as expected, is smaller than the one based on the naive estimate, given by 0.0426, calculated by (6.41). The PC model gives an average standard error of 0.0420, which underestimates the empirical value based on the Cox PH model 0.0434, only by 0.03%.

The results from this simulation show that the PC approximation to the Cox PH in simple cases like this, where only one parameter is estimated, can be very good. The selection of the pieces is crucial for this to occur, and as mentioned earlier, it is advised to determine the length and number of pieces based on the curvature of the baseline cumulative hazard function from the Cox PH model. Implementation of these methods on jobless spells from SLID is presented in the next chapter, where many variables are included in the initial models and the number of pieces has to be chosen carefully in order to avoid having too many parameters for estimation.

The implementation of the PC model approximation to a stratified Cox model as in (6.8) can be complicated; however, naive variance estimates that take the IPC weights as fixed, which are given by standard Cox PH software, can be used if they don't differ substantially from the random IPC based variance. Further development of the PC model approximation to stratified Cox PH models is needed.

Chapter 7

Implementation on SLID Data

In this chapter, jobless spells from SLID are used as an illustration of the methods and features of longitudinal survey data that have been described in preceding chapters. As mentioned earlier, each SLID panel is surveyed for six years, with annual interviews that collect labour and income information regarding the preceding year. We will focus in particular on jobless spells from the third panel of SLID, which was followed from 1999 to 2004. Because of the duration of the observation period and the widely spaced interviews, loss to follow-up is likely to be substantial in SLID. In the analysis of jobless spells from SLID, it seems natural to assume that individuals who experience longer jobless spells are more likely to drop out from the survey, than those who have shorter spells. Therefore, dependent loss to follow-up needs to be considered in the estimation of survival probability distributions and of regression coefficients.

The exploratory statistics presented in section 7.1 provide some of the general features of the SLID sample and also of a subset that has been determined considering various forms of loss to follow-up and features of the starting dates of the spells. Information supplementing the discussion is given in Appendix B.

Sections 7.2 and 7.3 regard the implementation of the estimation procedures described in chapters 5 and 6 concerning the estimation of survival probability distributions via the Kaplan-Meier estimate and also the analysis of covariate effects through the Cox PH model. Supplementing information on the Kaplan-Meier estimation is given in Appendix D.

7.1 Introduction

The target population of SLID is composed of all persons living in Canada, excluding those residing in Yukon, Northwest Territories or Nunavut, in institutions or in Indian reserves and full-time members of the Canadian Armed Forces living in barracks. As mentioned earlier, at the beginning of the survey, an initial sample is drawn from the Labour Force Survey (LFS), which is based on a stratified multi-stage sampling design. Each province is divided into LFS economic regions, which are then subdivided into one or more urban and rural areas. Further subdivision of the urban areas is based on socioeconomic characteristics. The primary sampling units (PSU's) are clusters formed of groups of dwellings within each stratum, and a random sample of households is finally selected within PSU's.

One of the most challenging issues arising from the analysis of jobless spells from SLID has been that of missing data, which is present in various forms. Some of them have to do with the start and end of the jobless spells, others with missing covariate information. The missing data problem is further discussed in chapter 8.

An issue regarding possible measurement error arises since the start and end dates of jobless spells are obtained through the employment information collected during the SLID interviews, rather than unemployment information. That is, the person that is interviewed is not asked to provide dates regarding their jobless spells, but those dates that are related to their past work experience. Then calculations are done in order to create SLID variables regarding the start and end dates of the jobless spells. In many cases the start dates of the jobless spells are coded with a "don't know" response making the length of the spell impossible to compute. The missing value in the start date variable has further impact in the modelling of dropout through covariates that are based on jobless spells characteristics.

The SLID information that refers to the termination of the jobless spells has also been an issue. There are spells in SLID that are not observed to completion. To account for this, there is a SLID variable "endtyp7" that provides information about the end of a jobless spell. If the spell was completely observed, endtyp7=1, otherwise endtyp7=2, either because (i) the respondent reported working in subsequent interviews, (ii) there was

non-response or (iii) the person was no longer eligible for the labour interview. Reason (i) is due to a recollection error, in which the person had reported being jobless at a particular time in one interview and in the next he or she reports the contrary for that same time. We have equated the two latter possibilities with dropout, but the question remains on what is the best way to deal with the former. This issue is further discussed in the appendix, section B.4.

Another type of missing information has been found in variables that are not directly related to the start and end of jobless spells, but are intended to be used as covariates for duration or dropout models. In SLID, this missing information comes in the form of responses of the types "don't know", "not available" or "refusal". As a way to deal with this, even though it is not ideal, missing values were included in most models as an additional covariate level, for the covariates that are categorical. The only continuous covariate was Age, which did not have missing values. The variable related to marital status in the models had only a few missing values and these cases were excluded from analysis. These missing values affected the estimation of the IPC weights for these individuals, because a person who had a missing marital status value in any of the six years of SLID would have an incomplete set of estimated dropout related probabilities to compute IPC weights.

SLID provides a variety of sampling weights to choose from. As will be mentioned later on, the right option is not always clear and depends on the objectives of the analysis. SLID weights account for several forms of attrition, are calibrated against population totals and are further adjusted for other factors. Since we are already accounting for non-response and attrition in our IPC weights, we prefer a set of sampling weights that are clear from these kinds of adjustments. SLID weights can be longitudinal or cross-sectional. In both cases, there is a set of weights available for each person and each year of the observation period of six years. Longitudinal weights are designed to represent the Canadian population from the year in which a panel started being interviewed. For instance, each year from 1999 to 2004, the longitudinal weights from panel 3 are adjusted for non-response and for influential weight values to represent the population from 1999. A weight value is considered influential if it has an excessive effect on the income estimate of total provincial income. After these adjustments, the weights are further calibrated

with respect to population totals in order to represent the population from 1999.

Cross-sectional weights are designed to represent the population of each year of a longitudinal panel. They are adjusted for the same features as the longitudinal weights, and also take into account interprovincial migration, the addition of cohabitants in the households, and panel allocation of each one of the six years. The panel allocation refers to the overlap of panels in a given year. For example, the samples from panels 2 (1996-2001) and 3 (1999-2004) are integrated by a single set of weights in order to have a larger sample for cross-sectional analysis separately for years 1999, 2000 and 2001. For details about the longitudinal and cross sectional weights construction, see LaRoche, [30].

The SLID sample of panel 3 involves the years 1999-2004 and will be the object of our attention in this chapter. It is composed of a total of 43683 individuals, of which 41% had at least one observed jobless spell in the six years. During the jobless periods, individuals may or may not have been looking for work. In many cases the jobless spells start a long time prior to the first interview in 1999. Furthermore, in our analysis we assume a monotone type of missing information, which means that the first time a person missed an interview is considered as the dropout time. Information that is available from subsequent interviews will not be used in our analyses. More discussion of loss to follow-up is contained in Appendix B1.

7.2 Some general features of jobless spells from SLID

A jobless spell in SLID is defined as the period of time in which a person is out of work and may or may not be looking for work. Note that this definition includes people who are out of the labour force (a person who does not actively look for work during a jobless spell is considered to be out of the labour force. See appendix B, section B.2). The jobless spells used for the exploratory statistics in section 7.1 pertain to this definition while the analyses in sections 7.2 and 7.3 involve jobless spells where the person was looking for work.

In the SLID longitudinal panels, non-response is represented by zero longitudinal

weights, assigned to people that did not respond to the interview, nor anybody in their household. These weights are positive if at least one person from the household responded to the interview. Response rates can be calculated by comparing the number of people that responded to the interview among the total number of individuals in the sample. Table 7.1 provides counts of the 43683 individuals from the SLID sample by response status, longitudinal weight and year. The response status is indicated by (01) if in scope (living in any of the 10 Canadian provinces), (02-06) if out of scope, and (07) if dropped out from the survey. The response rates in the last column are calculated by dividing the number of respondents (with w>0) by the number of longitudinal persons in the sample (43683). Note that about 28% of the people moved out of scope, were deceased or dropped out from the survey by 2004 (response status 02-07).

Table 7.1: Number of SLID individuals by response status, longitudinal weight (w) and year, SLID sample.

Year	Weight	Response status*				Subtotal	Total	Response rate
		01	02-06	07	08			
1999	w=0	7,024	-	-	-	7,024		
	w>0	36,158	501	-	-	36,659	43,683	0.84
2000	w=0	6,643	-	801	-	7,444		
	w>0	35,340	899	-	-	36,239	43,683	0.83
2001	w=0	4,292	-	3,147	-	7,439		
	w>0	34,892	1,352	-	-	36,244	43,683	0.83
2002	w=0	3,970	-	4,941	3	8,914		
	w>0	32,922	1,847	-	-	34,769	43,683	0.80
2003	w=0	2,661	-	7,625	9	10,295		
	w>0	31,214	2,174	-	-	33,388	43,683	0.76
2004	w=0	1,701	-	9,778	10	11,489		
	w>0	29,631	2,563	-	-	32,194	43,683	0.74

* Response status (SLID variable resp99 codes):
01=in scope; 02-04=living outside of the 10 Canadian provinces;
05=institutionalized; 06=deceased; 07=dropped out from survey;
08=not real person.

Since there are several ways in which attrition can occur in SLID, the convention was made for our purposes, that a person will be considered to be observed in a particular year (that is, not lost to follow-up) if their household had been reached and also if labour information on them was available in that year. Individuals who were not observed in the first two consecutive years of the panel were excluded from analysis. A person will be

considered lost to follow-up the first year in which they (i) were a non-respondent, (ii) had no job information, (iii) were out of scope, or (iv) dropped out from the sample. Please refer to section B.1 in the appendix for more detailed information about these SLID types of loss to follow-up (LTF) and the selection criteria. In our analysis we assume a monotone type of missing information, which means that the first time a person missed an interview is considered as their dropout time and so information regarding people and their jobless spells that is available from subsequent interviews will not be used in our analyses.

Table 7.2 involves Canadians who were at least 11 years old in 1999, who may or may not have experienced one or more jobless spells in 1999-2004. Part (a) shows the number of individuals and part (b) the corresponding number of jobless spells, while the second column involves the SLID sample and the third column involves individuals after adjusting for LTF (section B.1 in the appendix). The data set referred to in the third column of Table 7.2 can be used to model LTF. Table 7.3 is analogous to Table 7.1 and shows the number of people by longitudinal weight, response status and year for the 32834 individuals selected for LTF modelling. Note that after adjusting the for LTF, the percentage of people in the new data set that moved out of scope, were deceased or dropped out from the survey is about 21%.

Table 7.2: Counts of individuals and jobless spells from SLID before and after adjusting for LTF.

(a) No. individuals:

No. spells	SLID sample	After LTF adjustment
0	25,674	15,339
≥ 1	18,009	17,495
Total	43,683	32,834

(b) Corresponding no. spells:

Start date	SLID sample	After LTF adjustment
Known	23,256	22,914
Unknown	8,320	7,869
Total	31,576	30,783

The year in which a person was last seen corresponds to the year prior to which they were LTF. Table 7.4 shows the number of people that were last seen by year, from

Table 7.3: Number of SLID individuals by response status, longitudinal weight (w) and year, data set resulting after LTF adjustment.

		Response status*					
Year	Weight	01	02-06	07	Subtotal	Total	Response rate
1999	w=0	1,981	-	-	1,981		
	w>0	30,838	15	-	30,853	32,834	0.94
2000	w=0	1,986	-	436	2,422		
	w>0	30,082	330	-	30,412	32,834	0.93
2001	w=0	2,519	-	488	3,007		
	w>0	29,083	744	-	29,827	32,834	0.91
2002	w=0	2,669	-	1,593	4,262		
	w>0	27,367	1,205	-	28,572	32,834	0.87
2003	w=0	2,016	-	3,315	5,331		
	w>0	25,946	1,557	-	27,503	32,834	0.84
2004	w=0	1,383	-	4,895	6,278		
	w>0	24,602	1,954	-	26,556	32,834	0.81

* Response status (SLID variable resp99 codes):
01=in scope; 02-04=living outside of the 10 Canadian provinces;
05=institutionalized; 06=deceased; 07=dropped out from survey.

the data set that resulted after adjusting for LTF. Individuals that were observed in the second year but not in the first were included as if they had joined the sample in the second year (that is, in 2000), a total of 2372. The table also shows the yearly LTF rate with respect of the remaining individuals in the sample. Note that about 42% of the people were LTF by the end of the six years (based on the LTF definition in appendix, section B.1).

Since economic conditions may change year after year, while analyzing jobless spells it makes sense to exclude spells that started before January 1st of 1999 (about 26% of the spells, as shown in table B.2 in Appendix B). Furthermore, there is uncertainty about the accuracy of some start dates that refer to a long time before 1999, in some cases even of several decades. The number of jobless spells starting between 1999 and 2004 from individuals 16-69 years of age in 1999, is 20669 (5474 unknown), and they correspond to 11881 individuals who may or may not have been looking for work. The following tables concern to these 20669 jobless spells.

Table 7.5 shows counts of jobless spells by spell order as well as the number of spells with known and unknown start date. Note that most missing start dates correspond to

Table 7.4: Number of individuals last seen by year, data set resulting after LTF adjustment.

Year	Last Seen	LTF	Remaining in sample	Yearly LTF Rate
1	4,570	0	32,834	-
2	2,878	4,570	28,264	0.162
3	3,167	2,878	25,386	0.113
4	1,827	3,167	22,219	0.143
5	1,412	1,827	20,392	0.090
6	18,980	1,412	18,980	0.074

Note: 2372 persons started follow-up in year 2.

the first spells. Spells with unknown start date were discarded from subsequent analyses. Since this may incur potential bias, methods to deal with missing unknown start dates are needed, as is discussed in chapter 8.

Table 7.5: Number of spells with known and unknown start date by spell order, after adjusting for LTF, start date and age range. During these spells the person may or may not have looked for work.

	Start Date		
Spell order	Known	Unknown	Total
1	6,930	4,951	11,881
2	4,402	326	4,728
3	2,082	130	2,212
4	1,010	47	1,057
5 +	771	20	791
Total	15,195	5,474	20,669

The distribution by spell order and starting year is described in Table 7.6, among the 15195 spells with a known start date. Most of the first spells start in the years 1, 2, and 3 of the panel (corresponding to years 1999-2001). Second spells more frequently start in years 2 or 3, third spells in years 3 and 4, and so on. Also, most of the first

and second spells start in the first three years of the panel. Among spells with known start date, Table 7.7 shows the number of individuals who had 1,2, and up to 5+ spells. It shows that 74% of individuals had one or two spells, contrasting with the 81% in the initial SLID sample, as shown in Table B.3 in Appendix B. The spell frequency from an individual is likely to be related to the spells' length. For instance, those spells that belong to individuals who experienced them once in the six years may be lengthier than spells from persons who experienced four spells or more.

Table 7.6: Number of jobless spells by start year and order. During these spells the person may or may not have looked for work.

	Start Year*						
Spell Order	1	2	3	4	5	6	Total
1	2,810	1,442	1,062	654	538	424	6,930
2	486	1,064	985	746	612	509	4,402
3	33	230	509	524	398	388	2,082
4	<15	∼38	138	254	301	275	1,010
5+	0	<15	∼50	148	245	312	771
Total	∼3,340	∼2,783	∼2,744	2,326	2,094	1,908	15,195

* Corresponding to years 1999 - 2004.

∼ Approximate due to confidentiality.

Table 7.7: Number of individuals by number of spells (with a known start date, and in which the person may or may not have looked for work).

Spells	Individuals	Percent
1	4,100	46.88
2	2,434	27.83
3	1,154	13.20
4	590	6.75
5+	467	5.34
Total:	8,745	100.00

The number of censored and uncensored spells with a known start date are shown

in Table 7.8. The percent of censored spells for each order ranges from 28.3% to 35.1%. Censored spells from all orders are 31.7% of the total.

Table 7.8: Counts of censored and not censored spells by order of spell, among spells with known start date.

Spell order	Not Censored	Censored	Total	Percent Censored
1	4,547	2,379	6,930	34.3
2	3,155	1,247	4,402	28.3
3	1,465	617	2,082	29.6
4	708	302	1,010	29.9
5+	500	271	771	35.1
Total	10,375	4,816	15,195	31.7

It is important to note that because of the high degree of missing data it will not be possible to draw firm conclusions from the analysis of jobless spells presented in section 7.3. However, the analysis illustrates the methodology developed in this thesis. Chapter 8 provides further discussion regarding data quality issues and problems in using longitudinal survey data for inference about life history processes.

7.3 Loss to Follow-up Modelling in Ontario and Quebec

As mentioned earlier, a jobless spell in SLID is defined as the period of time in which a person is out of work and may or may not be looking for work. The jobless spells used for the exploratory statistics shown in section 7.1 pertain to this definition while the analyses in sections 7.2 and 7.3 involve jobless spells where the person was looking for work. For a detailed definition of jobless spells in SLID refer to the appendix, section B.2.

This section begins with a brief summary regarding the modelling of loss to follow-up, followed by the implementation of the weighted Kaplan-Meier and the Cox PH model-based analysis, discussed in chapters 5 and 6.

A person is considered lost to follow-up in a given year if he or she meets one of the following: (i) the person was a non-respondent, (ii) had no job information, (iii) was out of scope, or (iv) dropped out from the survey. Individuals were followed until the first year in which they experienced any of these four conditions, and may have started follow-up in the first or second years of the panel (more detailed information can be found in the Appendix, section B.1).

The logistic model discussed in section 4.1, in expression (4.5), was used to describe LTF from SLID. Models were fit for the years 2000 to 2004. The selection of covariates was based on the list of variables that are used for non-response modeling in SLID (see La Roche [30]). As will be described shortly, one covariate related to the jobless spells was added, which indicates whether the individual was jobless in the preceding interview. Recall that in SLID, non-response is one of the four conditions that define our dropout response variable (if nobody in the household responded to the interview, the person meets condition (i) above).

The sample composition evolves over the survey years, in the sense that it includes people that turn 16 years of age each year. We are analyzing individuals that are eligible to provide labour information, that is, are 16 years or older in a given year. Recall that covariates in the LTF model for year t are based on information to the end of year $t-1$. This means that a person who is analyzed for LTF in the year 2000 must have labour information in the year 1999. For this reason, LTF is analyzed for people not younger than 16 years in 1999. Furthermore, there was a considerable number of individuals with a code of "9:NA" (missing values) in the variable Student referring to whether the person was a part or full time student, or was not a student . It has been found that all individuals with this type of code in a given year were 69 years or older. The estimation of dropout probabilities (and consequently of spell durations) was therefore restricted to persons who were between 16 and 64 years of age in 1999.

The weighted Kaplan-Meier estimation in the next section is performed for jobless spells from residents of Ontario and Quebec, separately. Loss to follow-up was thus modelled for these two provinces. There is a set of individuals from whom labour information was available only starting in the year 2000 and they were included in the samples from Ontario and Quebec, starting from this year. The number of individuals in the sample

from Ontario is 4412, from which 40% were LTF by year six, including the 664 individuals that joined this sample in year 2000. The Quebec sample is of 3102, from which 43% were LTF by year six, including the 222 individuals that joined the sample in year 2000. Table 7.9 shows this and the number of individuals that were used in each one of the models for years 2000 to 2004. Detailed information about the variables and counts of individuals within each category can be found in Appendix C, Tables C.1 and C.2 for Ontario and Quebec, respectively. Note in Table 7.9 that the group of individuals that started follow-up in year 2000 are much more likely to drop out later on and may have a different covariate distribution than those that joined the sample in 1999. This issue was not dealt with in the dropout modelling here; however it would have been ideal to use a separate dropout model for these groups.

The names of the variables used in the LTF models are as follows: Sex, Age, Education Level ("Edlev"), Marital Status ("Marst"), Immigration Status ("Immst"), Student ("Stud"), Renter, Household Size ("HHsz"), Family Composition ("Famtype"), Household Type ("HHtype"), Urban and interactions. There is one variable that is directly related to the durations of jobless spells, recording whether an individual was jobless at the previous interview time, and is denoted by "Jstat". The logistic model from expression (4.5) was fitted for each year of the panel, starting from year 2000 to 2004 (or $t \in \{2, 3, 4, 5, 6\}$). The "Jstat" covariate was significant in all five models. Table 7.10 indicates with an "x" the variables that were significant at the 5% level for the final LTF models of each year in Ontario and Quebec, respectively.

The selection of covariates was carried out by backward elimination using SAS. Informal model checks were performed by implementing the methods from Hosmer and Lemeshow [24] regarding deciles of risk. These tests indicate that the models fit the data satisfactorily. A detailed discussion of the modeling construction and evaluation can be found in Appendix C.

Table 7.9: Counts of individuals for the LTF model by year and province.

	Ontario		Quebec	
	Joined since		Joined since	
LTF Model	1999	2000	1999	2000
2000	4412	0	3102	0
2001	3768	664	2557	222
2002	3412	468	2263	134
2003	2928	331	1972	97
2004	2746	278	1819	79

Table 7.10: Variables in final models by year, LTF model (Ontario and Quebec), years 2000-2004

	Ontario					Quebec				
Variable	2	3	4	5	6	2	3	4	5	6
Sex			x							
Age	x	x	x	x	x	x	x	x	x	x
Age^2	x	x	x	x	x	x	x	x	x	x
Edlev	x	x	x	x		x		x		x
Marst	x	x	x	x	x					
Immst	x	x					x		x	
Stud		x	x	x	x	x	x			
Renter	x	x	x							x
HHsz								x	x	
Famtype					x		x			
Hhtype				x			x			
Urban	x			x	x		x	x		
Jstat	x	x	x	x	x	x	x	x	x	x
Sex*Marst			x							
Age*Marst	x		x		x					
Urban*HHsz								x		

7.4 Kaplan Meier estimation of jobless duration distributions for residents of Ontario and Quebec

In this section we refer back to chapter 5, where weighted Kaplan-Meier (K-M) estimation is discussed. The estimates presented here are descriptive quantities of the jobless spells distributions from people living in Ontario and Quebec in the years 1999 and 2000. The finite population quantity to be estimated is the empirical distribution given in expression (5.1). As mentioned in section 5.1, it is useful to assume that the finite population quantity in (5.1) converges in probability to a superpopulation duration distribution $S(y)$ as the population's size increases to infinity, and so we can estimate the finite population quantity from a superpopulation approach. In the following discussion, the formula for the weighted Kaplan-Meier estimator in expression (5.9) and the estimation methods from section 5.2 are implemented. The variance estimates were computed as in expression (5.11) where $\widehat{Cov}[\hat{h}(s), \hat{h}(t)]$ is the (s,t) element of $\widehat{Var}(\hat{\theta})_{comb} = \hat{B}_{comb}^{-1}\hat{C}_{comb}\hat{B}_{comb}^{-1}$ as in expression (4.26), where $\hat{B}_{comb}$ and $\hat{C}_{comb}$ have the form in (4.27) and (4.28), respectively.

Jobless spells from individuals living in the provinces of Ontario and Quebec in 1999 were analyzed separately, with durations measured in weeks. The analyses shown here correspond to spells that started in 1999 and 2000. The strata and clusters that were used for variance estimation correspond to economic regions and dissemination areas from SLID. Recall that the economic regions are groups of census divisions which are intermediate geographic areas between the province and the municipality (census subdivision). The Dissemination Areas (DA's) are small areas composed of one or more neighbouring blocks and constitute the primary sampling units (PSU's). Each dissemination area is assigned a four digit code that is unique within a census division and a province or territory. In order to identify each DA uniquely in Canada, the two digit province code and the two digit census division code must precede the DA code.

There are $R_{ON} = 11$ and $R_{QC} = 17$ economic regions used as strata in the Ontario and Quebec samples. The number of clusters within provinces and start year that were used can be found in Table 7.11 below.

Among individuals that were used to model LTF from Ontario and Quebec in the preceding section, 30% and 36% had at least one jobless spell in the six years from 1999 to 2004. In total, spells from 1124 and 931 individuals living in Ontario and Quebec in 1999 were used to estimate the survivor function from the population. Table 7.11 also shows the number of complete and censored spells in each data set.

Table 7.11: No. of clusters, complete and censored spells by province and start year.

	Ontario			Quebec		
Start Year	No.clusters	No.spells	No.Cens.	No.clusters	No.spells	No.Cens.
1999	283	359	60	217	311	55
2000	220	270	30	162	211	21

Note: The clusters are Dissemination Areas, PSU's in SLID.

With the K-M estimates we aim to describe the jobless spells starting in 1999 and 2000 separately, from the population of individuals living in Ontario and Quebec in the year 1999. Having this in mind, one question arises regarding the right choice of sampling weights to use for estimation. It was mentioned earlier that in addition to the original longitudinal sampling weights associated with the start year of the panel interviews (year 1999), other weights are given at the end of every year from 2000 to 2004. Consider the example of a spell that started in 2000 and ended in 2001. It may be unclear which of the three available longitudinal weights should be used (1999, 2000, or 2001). One possibility is to use the weight corresponding to the end year of the spells. This option implies that weights from different years would be combined together in the analysis, which seems rather awkward. The weights from the start year, 2000, are an adjusted version of the 1999 weights, that takes into account for various forms of attrition and non-response. We must note though, that we don't require this adjusted version since our IPC weights are already adjusting for LTF. An advantage of using the original longitudinal weights from 1999 is that they are the closest to the base weights, ideally the ones to use for the analyses, but are not available from SLID. Cross-sectional weights from 1999 and 2000 given by SLID represent the populations from these two years. These would be appropriate if we were analyzing the jobless spells from people living in Ontario in these years. In our weighted Kaplan-Meier estimates, longitudinal weights from the year 1999

were used, since we want to describe jobless spells from individuals that were selected in the SLID sample in the year 1999.

Figures 7.1 and 7.2 show weighted Kaplan-Meier estimates from jobless spells experienced by residents of Ontario and Quebec in the year 1999. The two upper graphs are K-M estimates from spells that started in the years 1999 and 2000. The solid lines correspond to unweighted K-M estimates and the dashed and dotted lines represent the design and combined weighted estimates, respectively. Tables D.1 and D.2 in the appendix show a summary of the estimates.

The lower graphs in these figures show the corresponding standard errors based on several methods of variance estimation. One group, denoted in the graphs as DES and indicated by a solid line, consists of three different formulas that gave virtually identical values for all time points. The first method is the finite-population variance based on Binder [4] similar to expression (5.17), with the difference that the weights are sampling weights from SLID. The second method was based on Boudreau and Lawless [9], analogous to expression (5.15). The third one was based on Lin [38], where an extra term given by $\hat{B}^{-1}$ is added to the formula in (5.17). The dashed line corresponds to the combined (design $\times$ IPC) weighted method from chapter 5, denoted as COMB. The dash-dotted line represents the naive variance calculated using combined weights, but treating the IPC weights as fixed, analogous to expression (5.19) and denoted by COMB Naive.

These figures show that there is not a substantial difference between the weighted Kaplan-Meier estimates from the unweighted, DES, or COMB methods for Ontario in 1999; however, for the year 2000, the graph shows some difference, however, still small, in durations between 25 and 55 weeks, where the COMB method gives slightly higher estimates than DES. Estimates from Quebec in the year 1999 give very similar results between DES and COMB, with the former giving slightly higher values at the tail of the distribution, for durations of 55 weeks and longer, which accounts for 18% of the spells, approximately. The year 2000 estimate for QC also gives slightly higher values from the DES method, for durations of 37 weeks and longer, which have around a 0.38 probability.

Regarding the standard errors (lower graphs in the figures), it can be seen that the COMB Naive and DES can be very similar, as in the case of Ontario spells from 1999. The standard error based on the COMB Naive variance gives higher values than the other

two in the case of Ontario, and in the case of Quebec, larger variances are given by the DES method. For both provinces and years 1999 and 2000, the COMB standard errors give the smallest values. There is an increment in the values of the standard errors from the years 1999 to 2000 for both provinces.

The simulations from chapter 5 regarding Kaplan-Meier estimation show that the COMB method gave slightly smaller standard errors than the COMB Naive method. For jobless spells from SLID, this relation is preserved, however, the difference between these two methods is much greater. Looking back at the way in which the variance estimates are constructed, recall that the expression in (4.8) is equivalent to our variance formula (4.9). Note that the term $B_{ij}(\hat{\theta}, \hat{\alpha})$ in (4.8) corresponds to $Var\big(U(\hat{\theta}, \hat{\alpha})\big)$ when $i = j = 1$, and is related to $Var\big(G(\hat{\alpha})\big)$ and $Cov\big(U(\hat{\theta}, \hat{\alpha}), G(\hat{\alpha})\big)$ when $i = j = 2$ and $i = 1, j = 2$. The Naive variance results from $Cov\big(U(\hat{\theta}, \hat{\alpha}), G(\hat{\alpha})\big) = 0$, and as this covariance value increases, the greater the difference between COMB and Naive. The data from SLID has greater variability than simulated data, and many more covariates are used in the LTF models. This is reflected in the difference between the two methods.

Confidence intervals at a 95% confidence level were computed for the median, based on each one of the Kaplan-Meier estimates. This was done by finding a set of y-values satisfying $-1.96 \leq Z \leq 1.96$ where

$$Z = \frac{\hat{S}(y) - 0.5}{\mathrm{se}\big(\hat{S}(y)\big)}.$$

Since Z changes only at the observed duration times, we took the y values at which Z changes from being outside $(-1.96, 1.96)$ to inside $(-1.96, 1.96)$ (Lawless [32], p. 93).

Table 7.12 shows the estimated median and a 95% confidence interval, from the jobless spells distribution from years 1999 and 2000 and both provinces. The DES and COMB methods do not differ greatly in the province of Quebec, the median values are the same and the confidence intervals vary by one week. In the case of Ontario, the DES method gives slightly smaller estimated medians and confidence intervals than COMB.

Conclusions based on the COMB method are that spells decrease in median length from year 1999 to year 2000, for about 5 and 4 weeks for Ontario and Quebec, respectively. The confidence intervals from the Quebec spells are wider, reflecting greater variability. The median length of the spells in Ontario is greater than in Quebec for both years, the

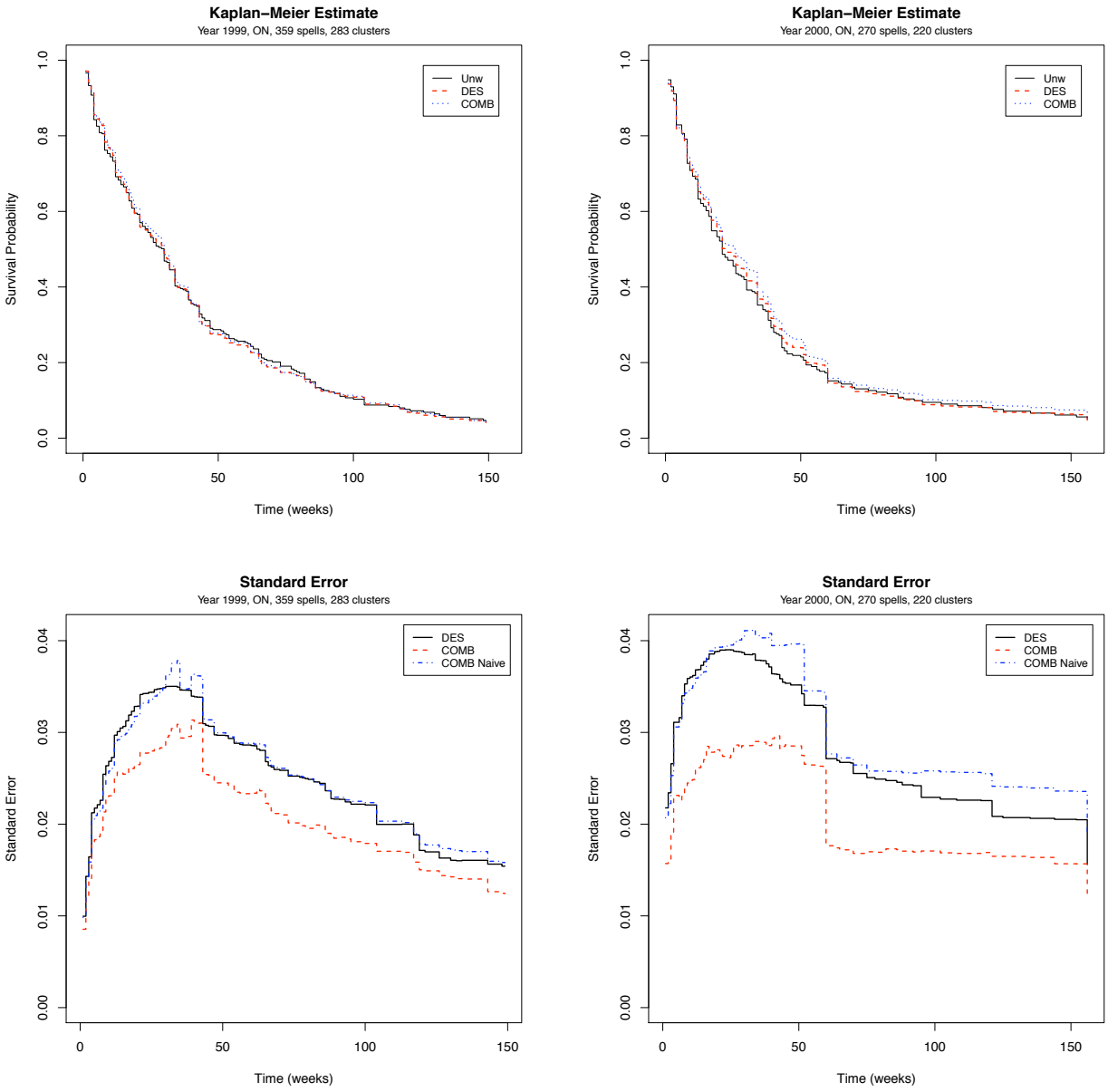

Figure 7.1: Weighted K-M estimates and point-wise standard errors from jobless spells starting in 1999 and 2000, from people living in Ontario in 1999.

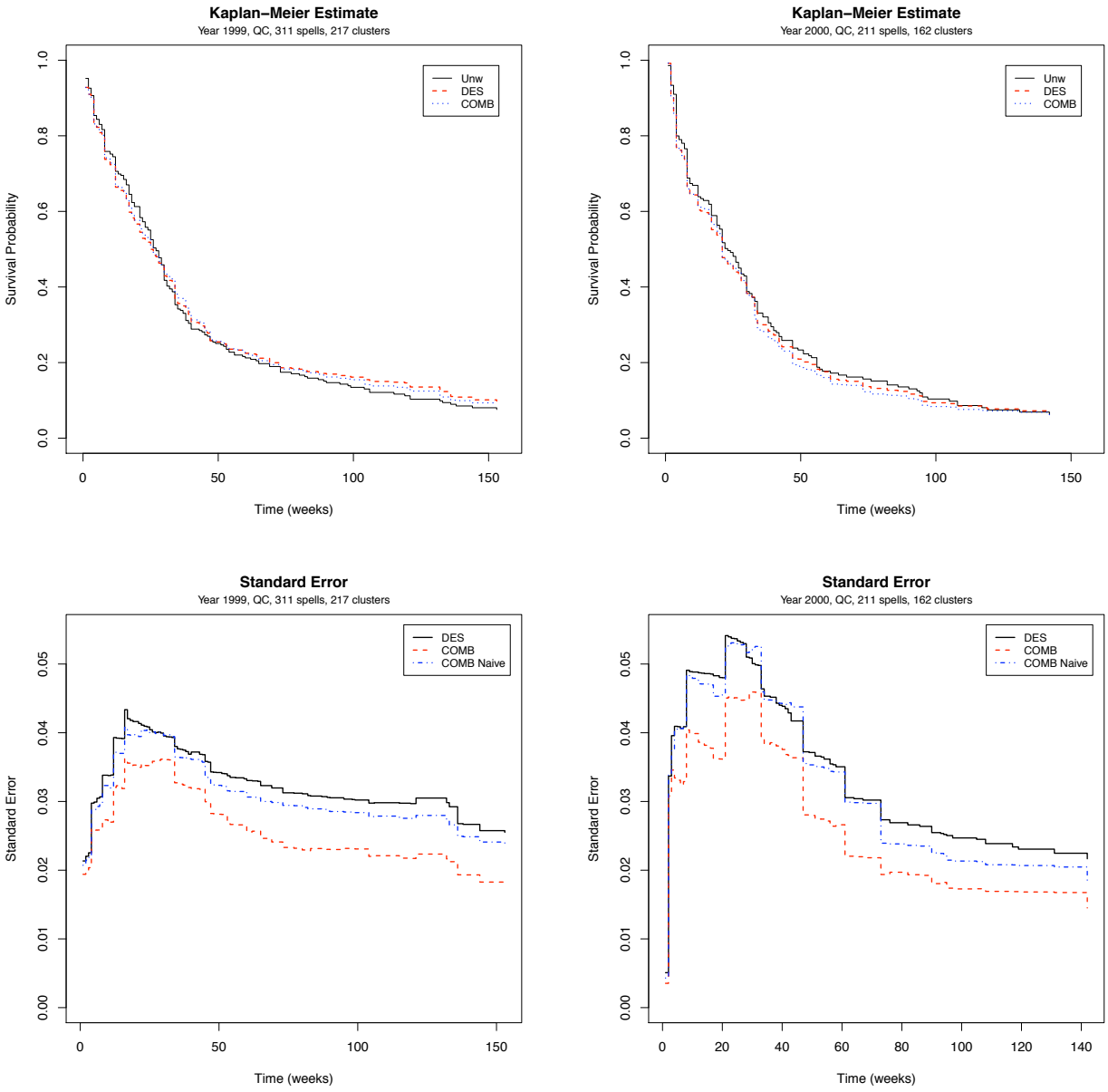

Figure 7.2: Weighted K-M estimates and point-wise standard errors from jobless spells starting in 1999 and 2000, from people living in Quebec in 1999.

medians differ by 5 and 4 weeks for 1999 and 2000.

Since spells with missing start dates were discarded and because of the great amount of missing data in SLID, these conclusions should be interpreted with care, and viewed as part of the illustration of the implementation of the variance estimates for IPC weighted Kaplan-Meier estimates.

Table 7.12: Estimated median for survival and 95% CI, Ontario and Quebec.

		Ontario		Quebec	
Year	Method	Median	CI	Median	CI
1999	DES	29	(21, 32)	25	(19, 31)
2000	DES	21	(19, 29)	21	(16, 29)
1999	COMB	30	(25,32)	25	(19, 30)
2000	COMB	25	(21, 30)	21	(17, 29)

7.5 Cox PH analysis of jobless durations for Ontario residents

In this section we give two analyses on jobless spells from individuals that were living in Ontario in the year 1999 and that were 16 to 64 years old in this year. The first analysis involves single jobless spells that started in the year 2000 and that were the first observed spells since 1999. The second analysis considers sequences of jobless spells that started in the years from 2000 to 2002.

The idea of presenting these two examples is to illustrate cases where interest lies in examining single spells and also cases where it is wished to examine sequences of spells in a specific period. In the former case, we account for information regarding job experience in the year preceding the jobless spells, 1999. In the case of sequences of spells, we include information prior to the start year of the spells and also information involving the sequences themselves, by adding covariates that refer to whether the spell is the first of a sequence and, if not, the length of the preceding spell. The IPC weights that were used are based on the LTF models presented in section 7.3.1 and Appendix C.

The analyses start by giving an assessment of the Piece-wise Constant (PC) model approximation to the Cox PH model. The approximation of the variance estimates is done by comparing standard errors provided by the `coxph` function in R/SPlus with those computed from the PC model. Variance estimates from design and combined methods that have been referred to in chapter 6 are also compared. Another part of the analyses presents a model obtained by variable selection, based on results from the combined weighted method. It should be noted that because of the missing data issues in SLID the results and conclusions from final models should be taken only as an illustration.

The PC model approximation to stratified Cox PH models was not implemented here; however, it remains of interest in this analysis to compare unstratified and stratified regression coefficients from the Cox PH model. Stratification is given as before, by the economic regions within Ontario (a total of $R = 11$); and clustering is given by the dissemination areas from SLID: 181 and 389 for the first and second analysis, respectively.

The model checking techniques that are available in Cox PH model software that allow for case weights were used. Even though these are based on naive variance estimates,

they can still provide an indication of departures from model assumptions. The proportional hazards assumption was checked by using the technique of model expansion, which consists of adding a time dependent coefficient term interacting with each covariate in the model and performing one significance test at a time or by performing a global test on all coefficients (see Lawless, p. 361 [32], Therneau and Grambsch p. 130 [57]). The linearity assumption of the PH model was checked by adding an extra term accounting for the square of the variable Age, which is the only continuous variable in the models. Graphical residual checks were done to complement the analyses, but are not shown here for confidentiality reasons.

7.5.1 First jobless spells starting in 2000

The variables that were used as covariates in the models for this subsection are described in Table 7.13, numbers 1-6. Some of these variables were selected based on the analysis on jobless spells from SLID discussed in Kovacevic and Roberts [29]. The Age variable represents the age in years and corresponds to the year 1999, as well as the variables employment insurance, occupation and income. The baseline level refers to the lowest category in all covariates.

PC model approximation to Cox PH

Estimation was performed on single spells from 196 individuals of which 19 were censored. The intervals that were chosen for the PC model and the hazard estimates are shown in Table D.3 in Appendix D. There are nine pieces, which were determined by visually examining the curvature of the cumulative baseline hazard function from the Cox PH model. Figure 7.3 shows the weighted Kaplan-Meier estimate with a naive variance-based 95% confidence interval. Note that the tail of the distribution shows jobless spells durations of up to 200 weeks (about 3.85 years) in length. The Kaplan-Meier estimate shows that the median baseline duration is approximately 26 weeks, durations longer than one year have a 0.26 probability and durations longer than 2 years have a probability of about 0.08. The estimated baseline cumulative hazard function for the main effects model in Table 7.13 is shown in Figure 7.4, using combined weights. We can see by the graph

Table 7.13: Variables used in models for first jobless spells that started in 2000 (1-6) and sequences of jobless spells starting in 2000-2002 (1-9) from residents in Ontario in 1999.

No.	Variable	Level	Description
1	Age	Cont.	Age, centered*
2	Sex	1	Female
		2	Male
3	Minority group	1	Yes
		2	No
4	Employment insurance	1	No
		2	Yes
5	Occupation	1	Trades, transport, equipment operators
		2	Management, business, finance, administrative occupations
		3	Natural, applied sciences, health, social science, education, art, culture, sport
		4	Primary industry
		5	Processing, manufacturing, utilities
		6	Sales and service
		99	Missing values
6	Income	1	≤First quartile
		2	(First quartile,third quartile]
		3	>Third quartile
7	Yearly quarter	1	Spell onset Jan-Mar
		2	Spell onset Apr-Jun
		3	Spell onset Jul-Sep
		4	Spell onset Oct-Dec
8	Order	1	Order of jobless spell =1 (since 1999)
		2	Order of jobless spell >1 (since 1999)
9	P.Dur	Cont.	Length of previous jobless spell duration**

*Mean age in data set for analysis 1 is 34.41, in data for analysis 2 is 34.15

**Continuous, when Order=2 the mean value of 25.8 weeks.

that the estimates of the cumulative hazards based on the Cox PH are well approximated by those based on the PC model.

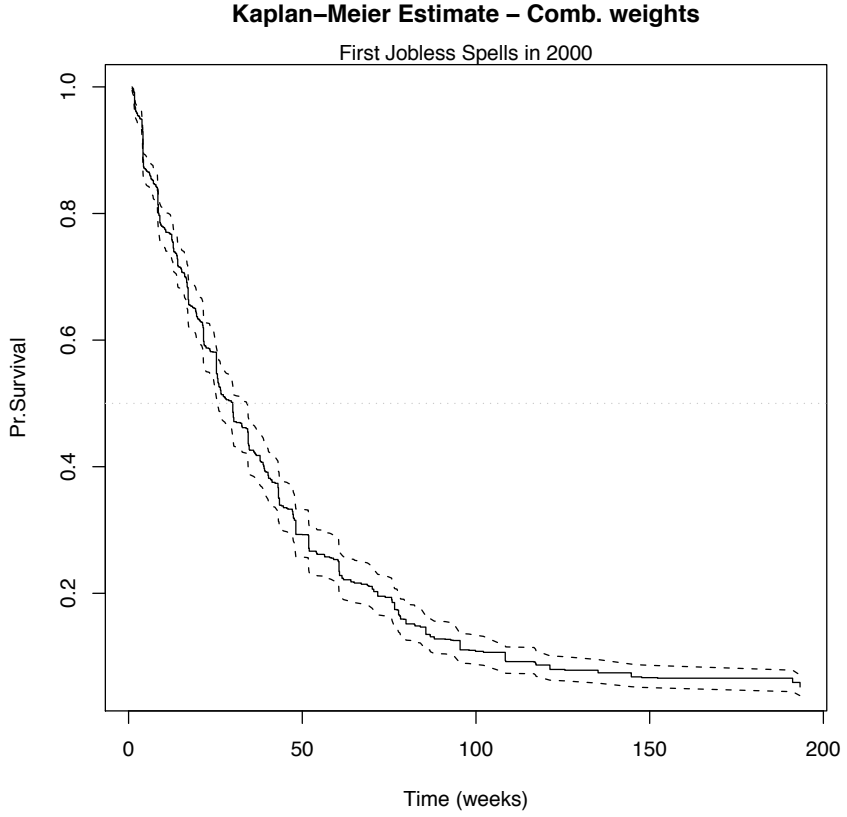

Figure 7.3: Weighted Kaplan-Meier based on the main effects model from first jobless spells in 2000, using combined weights.

Estimates of the baseline cumulative hazard functions for the main effects Cox PH regression model and of the regression coefficients were obtained using the `coxph` function in R/SPlus. Estimates based on the PC model were computed according to the methods described in chapter 6, with variance estimates that account for 181 clusters and 11 strata. The values of unweighted and weighted estimates are shown Tables D.4 and D.5 in the appendix, respectively.

The PC model gives estimated regression coefficients that are mostly higher than those from the Cox PH model. Tables D.4 and D.5 show that the unweighted, design and combined weighted regression estimates over-approximate the Cox PH estimates

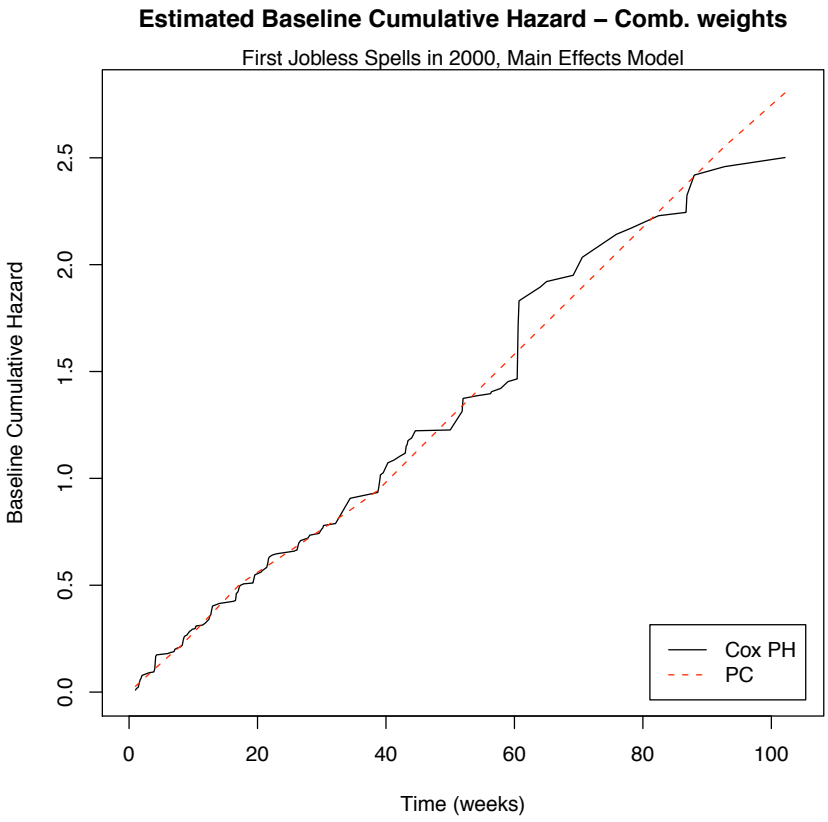

Figure 7.4: Weighted estimated baseline cumulative hazard function for the main effects model from first jobless spells in 2000, using combined weights.

by an average of 9%, 10% and 6%, respectively. The variance estimates from the Cox PH that were used for comparison are robust variance estimates from the `coxph` and `cluster` options in R/SPlus, based on Lin and Wei [40] (do not include stratification). The PC model variance analogues give in the unweighted case, an average of 3% of over approximation, while the design and combined method give an average of 4% and 5%, respectively.

The graphs in Figure 7.5 show z-statistics $|\hat{\beta}^c|/se(\hat{\beta})$ where the numerator is based on point estimates from the Cox PH model and the denominator, on variance estimates from the Cox PH or the PC models. The left graph shows estimates based on sampling design weights and the right graph on combined (design×IPC) weights. The z-values based on robust variance estimation methods are labeled as "Cox PH.robust" and "PC.robust", while those involving variance estimates from the PC model, based on Boudreau and Lawless [9], are labeled as "PC.B&L". Variance estimation methods analogous to those from Binder [4] and Lin [38] based on the PC model gave very similar results to those from PC.B&L, and therefore have been omitted.

The Cox PH.robust z-values are similar to those from the PC.robust method in both graphs, indicating a good PC approximation in terms of variance estimates. Design and combined methods agree that the three most significant variables are Age, Occup99 and Income.cat3; however, the combined PC method produces higher z-values, therefore smaller variance estimates, than the other methods. The remaining variables are less significant for the combined than for the design methods, therefore in this particular case, it seems that the combined methods separate highly significant from not so highly significant variables in a more evident way.

As expected, the standard errors based on the PC.Comb.Naive method are larger than those from PC.Comb, but without drastically changing the results in terms of significance. No stratification effects are indicated by the similarity of variances between the CoxPH.robust and PC.robust to PC.B&L in the design case and to PC.Comb.Naive in combined case.

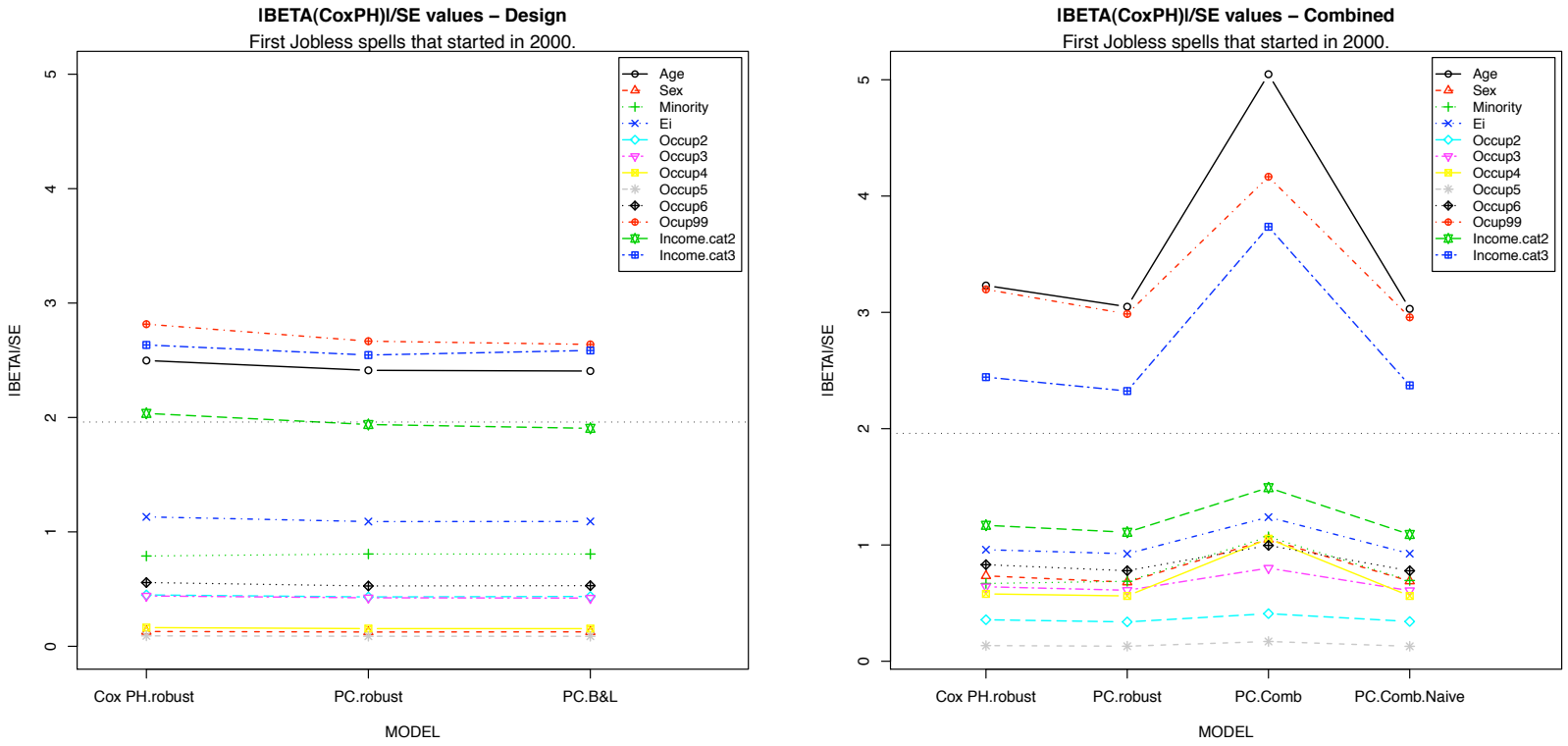

Figure 7.5: Values of $|\hat{\beta}^c|/se(\hat{\beta})$ from first jobless spells starting in 2000.

Model selection and interpretation of regression estimates

The model selection process consisted of a backward elimination procedure, based on the combined weighted methods from chapter 6. The main effects model with variables 1-6 in Table 7.13 was evaluated, including the interactions Age:Sex, Minority:Income, and Age:Occupation.

A summary of results from fitting the final model is shown in Table 7.14. This table shows the standard errors that were obtained from the "PC.Comb" and from "Cox PH.robust" discussed in the preceding subsection. Recall that the former variance estimates take the IPC weights as random while the former are considered naive since takes the weights as fixed. The variables that are significant at a 10% level or less are shown in bold.

The Cox PH.robust variance estimates shown in Table 7.14 gave similar values to the naive variances based on the PC model (not shown), the latter including stratification. As observed in the preceding subsection, these variance estimates give larger values than the PC.Comb method and consequently, have the effect of reducing the values of the z-statistics. The variables that have a substantial variance increment are Ei, Income.cat3, and the interactions Minority:Ei, Age:Occup4 and Age:Occup5.

A positive regression coefficient of a covariate category A implies that individuals in this category have an increasing risk of leaving the "jobless" state, relative to individuals that belong to a reference category B. This means that individuals from category A will more likely experience shorter jobless spells than those from category B. Conversely, a negative coefficient implies a decreasing risk of leaving the "jobless" state, thus individuals from category A are more susceptible of experiencing longer spells than the category B individuals.

The negative age estimated coefficient implies that as age increases, the risk of leaving the jobless state decreases (which is equivalent of having a higher risk of experiencing longer spells). Moreover, the squared value of Age indicates that there is a non-linear relationship of this variable with the log-hazard function of the length of jobless spells. The risk of leaving the jobless state with respect to the age variable will decrease as age increases, at a slightly decreasing rate. For example, the risk is of 0.493 for individuals

of age 25, and it decreases to 0 for individuals of age 34.4 (which is the age mean value), and will further decrease to -0.316 for individuals of age 43.8.

Employment insurance is shown to be an important factor for the length of jobless spells. Its negative sign indicates that a jobless spell from an individual who had employment insurance in the preceding year, is more likely to be longer than a spell from a person who did not receive employment insurance in the preceding year. The positive interaction term with minority indicates that the risk of experiencing longer spells will be shorter if the person is not from a minority group. The risk of leaving the jobless state when receiving Ei the year before while not being from a visible minority group is $\exp(-.109) = 0.89$ which decreases when being a from a visible minority to $\exp(-1.745) = 0.1746$.

Positive significant income coefficients for both categories of this variable imply that people with lowest income tend to experience longer spells. The risk of having shorter spells for individuals in the highest income category is $\exp(0.79) = 2.2$ times greater than for those in the lowest category.

The occupation variable resulted with a Wald based p-value much smaller than 0.0001, and the individual coefficients show that the only level that is significant is Occup99, referring to the "missing-values" category. Its interaction with age is significant at a 5% level for types of occupation 4 (primary industry) and 5 (processing, manufacturing, utilities) and a 10% significant for occupation type 6 (sales and service). For occupation types 5 and 6, this positive interaction means that shorter spells are more likely for members of this occupation type, compared with members of the type 1(trades, transports, equipment operators), as their age increases. Conversely, in the case of occupation of type 4, as age increases, it is more likely to experience longer jobless spells compared to people in the occupation of type 1.

Model checks

As mentioned at the beginning of this section, it is possible to perform informal model checks based on output from Cox PH software. Even though these do not consider the IPC weights as random, they can still provide indication of departures from model assumptions. The proportional hazards assumption was checked by using the technique of model expansion, which consists of adding a time dependent coefficient term associated

Table 7.14: Summary from fitting unstratified Cox PH model to first jobless spells that started in 2000 from residents of Ontario in 1999.

		PC.Comb		Cox PH.Robust (Naive)	
Variable	Est.	SE	p-Val	SE	p-Val
Age	-0.043	0.023	**0.062**	0.027	0.110
Age2	0.001	6.E-04	**0.076**	7.E-04	0.180
Minority	0.021	0.188	0.909	0.279	0.940
Ei	-1.745	0.526	**0.001**	1.119	0.120
Occup2	0.102	0.284	0.719	0.317	0.750
Occup3	-0.300	0.291	0.302	0.371	0.420
Occup4	-0.142	0.435	0.745	0.581	0.810
Occup5	0.076	0.275	0.783	0.352	0.830
Occup6	-0.206	0.261	0.429	0.311	0.510
Occup99	-1.867	0.417	**<0.001**	0.541	**0.001**
Income.cat2	0.428	0.224	**0.056**	0.275	0.120
Income.cat3	0.790	0.245	**0.001**	0.324	**0.015**
Minority:Ei	1.615	0.601	**0.007**	1.162	0.160
Age:Occup2	0.018	0.026	0.474	0.030	0.540
Age:Occup3	0.014	0.027	0.611	0.033	0.680
Age:Occup4	-0.099	0.032	**0.002**	0.049	**0.044**
Age:Occup5	0.061	0.029	**0.033**	0.040	0.130
Age:Occup6	0.044	0.024	**0.061**	0.028	0.110
Age:Occup99	-0.046	0.029	0.108	0.036	0.200

Significant at a 10% level or less shown in bold.

with each covariate in the model and performing one significance test at a time or by performing a global test on all coefficients (see Lawless, p. 361 [32]). Each of these tests can be seen as a trend test applied to the relationship of the Schoenfeld residuals and time, or a function of time.

Table 7.15 shows the PH assessment results from fitting the model from Table 7.14 and using the R/SPlus function `cox.zph`. The column labeled as "rho" is the Pearson correlation between the scaled Schoenfeld residuals and time or a function of time, the column "χ^2" gives the corresponding test statistic, followed by the p-values in "p-Val". A detailed explanation on how these tests are constructed can be found in Therneau and Grambsch [57], p.130. The PH assessment was performed using the Kaplan-Meier transformation of time, which is less sensitive to censoring patterns than the identity or logarithmic transforms. Table 7.15 shows no significant departures from the PH assumption in the global test, although the variables Age2, Ei, Occup3, Occup99 and Minority:Ei do have a 10% level significant departure. Further examination of Schoenfeld residual plots (omitted for confidentiality) have led to conclude that the model coefficients do not vary dramatically with respect to time.

One way to deal with departures from the PH assumption involves the addition of a time dependent covariate in the model and testing for its significance; however this strategy was not pursued here.

Stratification

In the introduction of this section, it was mentioned that the PC model approximation to stratified Cox PH models has not yet been implemented, however, it remains of interest to compare stratified vs. unstratified estimates of the Cox PH regression coefficients. Stratification is given by the economic regions within Ontario, a total of $R = 11$.

Table 7.16 shows the results from fitting a stratified version of the model with variables in Table 7.14. The standard errors are based on the variance estimate from Boudreau and Lawless [9], and therefore naive regarding the random nature of the IPC weights. Figure 7.6 shows a graphical representation of the z-values $|\hat{\beta}|/se(\hat{\beta})$ based on the naive unstratified and the naive stratified estimates from Tables 7.14 and 7.16. Most z-values decreased after stratification, some important changes are observed for the Age:Occup2

Table 7.15: PH Assessment, Cox PH unstratified fit on first jobless spells in 2000.

Variable	rho	χ^2	p-Val
Age	-0.006	0.008	0.931
Age2	0.085	3.643	**0.056**
Minority	-0.023	0.137	0.711
Ei	-0.074	4.810	**0.028**
Occup2	-0.012	0.031	0.859
Occup3	-0.111	3.494	**0.062**
Occup4	-0.064	0.844	0.358
Occup5	0.025	0.184	0.668
Occup6	-0.015	0.048	0.827
Occup99	-0.110	5.897	**0.015**
Income.cat2	0.026	0.221	0.638
Income.cat3	0.051	1.131	0.288
Minority:Ei	0.061	3.102	**0.078**
Age:Occup2	-0.016	0.062	0.804
Age:Occup3	-0.051	0.675	0.412
Age:Occup4	-0.081	1.195	0.274
Age:Occup5	0.045	1.049	0.306
Age:Occup6	0.003	0.002	0.963
Age:Occup99	-0.096	2.817	**0.093**
GLOBAL	NA	24.762	0.169

Variables with significance of 10% or less are shown in bold.

variable (from 0.72 to 0.022) and Occup6 (from 0.79 to 0.12), both variables remaining not significant. The significance of Age:Occup4 increased considerably after stratifying, going from about 5% to a 1% significance, this was the only significant variable that had a major impact from stratification.

The PH tests for variables that appear significant in this stratified model do not show important departures from the PH assumption, the global measure however shows a larger value of the test statistic, compared to the unstratified model (Table 7.16).

It remains of interest to perform an analysis using stratified point and variance estimates taking the IPC weights as random. From the comparison between these two models using naive variance estimates, it is expected that the stratification results using the PC variance approximation will behave similarly, that is, with no dramatic changes in the z-values for most variables. It remains unclear if the stratified model offers better results in terms of the PH assumption in this example, and further assessment is required by accounting for the IPC weights as random.

From the results presented in this subsection, it can be concluded that based on these data, PC model variance estimates approximate quite well those from the Cox PH model, considering the sample size and the resulting large standard errors, and can be used for inference. In this example, the PC.Comb.Naive method gives very similar variance estimates to the robust variance Cox PH estimates, indicating that there is not an substantial effect of stratification.

It is important to note that, because of the considerable amount of missing information, the results from the analysis of jobless spells is offered here only as an illustration. The variables that were found the most significant for the durations of first jobless spells in 2000 for individuals living in Ontario in 1999, are age, income, employment insurance, employment insurance and the interactions Age:Occupation and Minority:Ei. Since there was found no important difference between unstratified and stratified models under naive variance estimates, and given that stratification has not played an important role in these data, it is inferred that a comparison using PC.Comb variance estimates would not have led to contrasting conclusions and that the significance results would have been preserved for most variables.

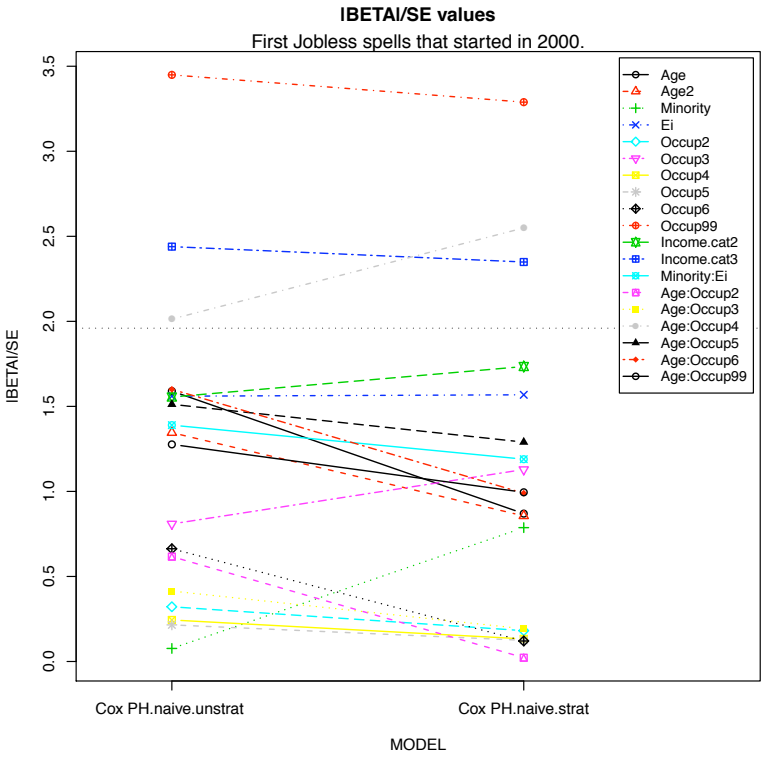

Figure 7.6: Values of $|\hat{\beta}|/se(\hat{\beta})$ from first jobless spells starting in 2000, for stratified and unstratified models, based on naive variance estimates.

Table 7.16: Summary of stratified fit and PH Assessment, first jobless spells in 2000, ON.

				PH Assessment		
Variable	Est.	SE	p-Val	rho	χ^2	p-Val
Age	0.001	0.001	0.380	0.085	1.986	0.159
Age2	-0.027	0.032	0.390	0.089	4.434	**0.035**
Minority	0.253	0.321	0.430	0.039	0.408	0.523
Ei	-1.778	1.134	0.120	-0.074	5.099	**0.024**
Occup2	0.071	0.390	0.860	0.044	0.526	0.468
Occup3	-0.531	0.470	0.260	-0.086	2.785	**0.095**
Occup4	0.093	0.704	0.890	-0.052	0.602	0.438
Occup5	0.052	0.421	0.900	0.035	0.401	0.526
Occup6	-0.045	0.374	0.900	0.062	1.073	0.300
Occup99	-1.935	0.588	**0.001**	-0.064	2.333	0.127
Income.cat2	0.441	0.254	**0.083**	0.058	1.375	0.241
Income.cat3	0.779	0.332	**0.019**	0.060	1.846	0.174
Minority:Ei	1.391	1.169	0.230	0.047	1.870	0.172
Age:Occup2	-0.001	0.035	0.980	-0.093	2.664	0.103
Age:Occup3	-0.008	0.041	0.850	-0.116	4.873	**0.027**
Age:Occup4	-0.150	0.059	**0.011**	-0.102	1.547	0.214
Age:Occup5	0.052	0.041	0.200	-0.005	0.011	0.916
Age:Occup6	0.032	0.032	0.320	-0.086	1.934	0.164
Age:Occup99	-0.041	0.041	0.320	-0.098	3.355	**0.067**
(GLOBAL)	-	-	-	NA	34.813	0.015

Variables significant at a 10% level or less are shown in bold.

7.5.2 Sequences of jobless spells in 2000-2002

This section presents an analysis of sequences of jobless spells that started in the period from 2000 to 2002, from individuals that lived in Ontario in 1999. To clarify, this means that some sequences had their first spell in this period and some started back in 1999. When analyzing sequences of spells, it is advisable to include information related to previous spells in the model in order to account for dependence on past event history.

This section is divided into several parts. As in the preceding analysis, the first one is about the assessment of the PC model approximation based on an initial model, the second part gives a discussion on a model selected by backward elimination, and is followed by model checks and stratified Cox PH modelling. We further include a brief discussion to illustrate the modelling of jobless spells by order, in a similar fashion as in Kovacevic and Roberts [29].

The variables that were used in the models are described in Table 7.13, numbers 1 to 6, are the same that were used in the preceding analysis, from which employment insurance, occupation and income refer to the year prior to the start of the spell; and three more variables (7 to 9) denoted as Yearly quarter, Order and P.Dur. The Yearly quarter variable refers to the quarter of the year in which the spells started and will be used to examine if the time of the year is an important factor in the length of the spells. The Order variable is equal to one if there was a previous spell and zero otherwise. This variable will be useful in finding out if second and subsequent spells are more likely to be shorter or longer than first spells. The length of the preceding jobless spell, given that there exists one, is denoted by P.Dur.

PC model approximation to Cox PH

In our sample of sequences of jobless spells in 2000-2002 from Ontario residents in 1999, there were initially 520 individuals with 655 spells; 61 spells were lost due to missing information in the variable P.Dur mentioned above, due to spells with unknown start date. Hence there are left 471 individuals with 594 spells. From these spells, 80 were censored, and 250, 206 and 138 spells started in years 2000, 2001, 2002.

The pieces for the PC model that were chosen and the estimated hazards are shown

in Table D.3 in the appendix, section D.2. There are nine pieces in total, and were determined based on a visual assessment of the curvature of the cumulative baseline hazard function from the Cox PH model which is very similar to the one estimated from the first jobless spells that started in year 2000, in the preceding analysis, and the same limits for the pieces were used.

The left hand side of Figure 7.7 shows a plot of the estimated survivor function and the right hand side is the estimated cumulative hazard function for the main effects model, based on the PC and the Cox PH models, respectively. The former approximates fairly well the estimate based on the latter. This plot also shows that the slope becomes lower after 52 weeks, indicating a decreasing hazard rate after this time, that is, the conditional probability of leaving the jobless state after 52 weeks decreases with time.

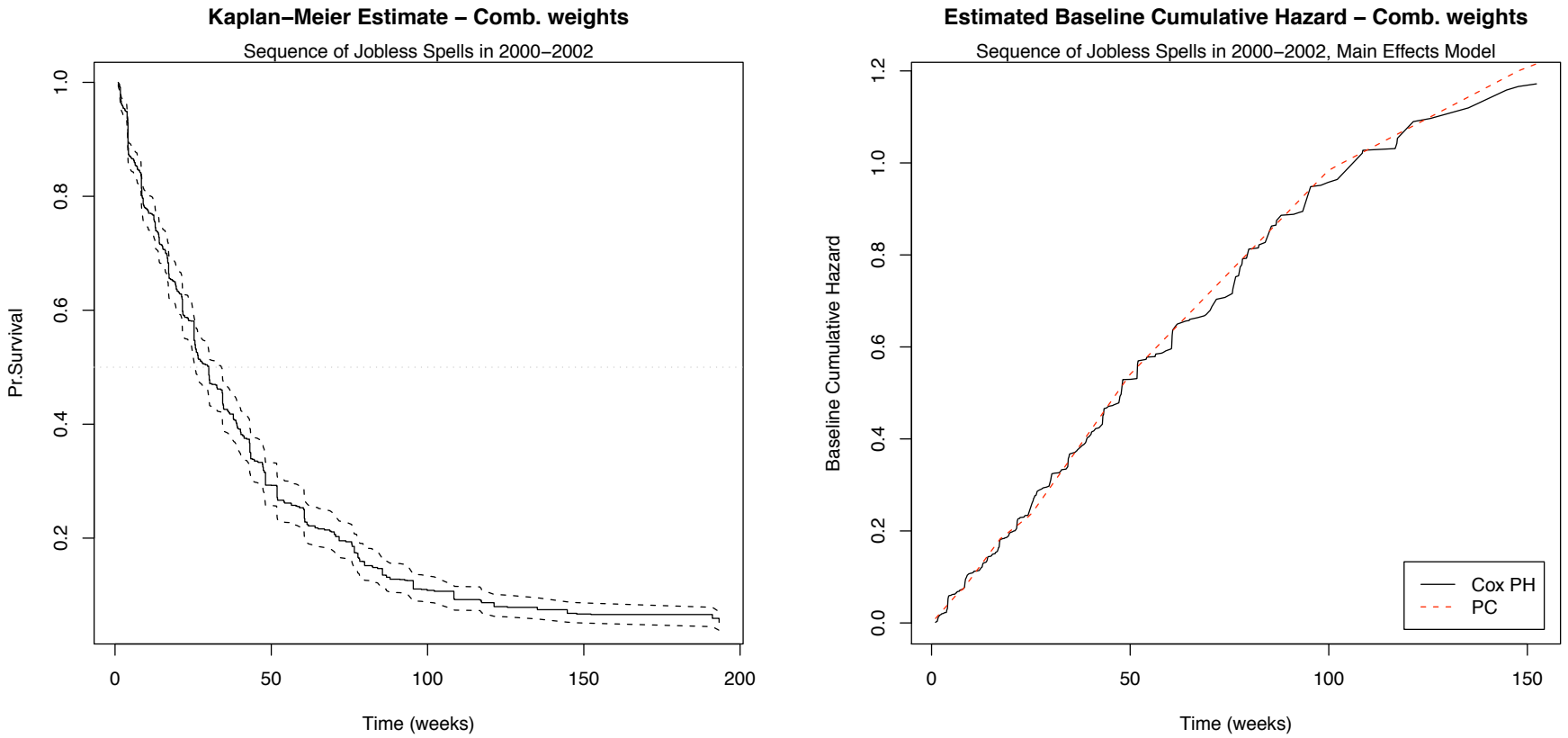

Figure 7.7: Weighted Kaplan-Meier and estimated baseline cumulative hazard function for main effects model from sequences of jobless spells in 2000-2002, using combined weights.

Estimates of the regression coefficients based on the Cox PH model were obtained using the `coxph` function in R/SPlus, and those based on the PC model were computed according to the methods described in chapter 6, with variance estimates that account for 389 clusters and 11 strata.

The estimation results are shown in Tables D.6 to D.8 in the appendix. Both PC model based estimates are calculated based on a larger sample size than those from the preceding section, and do a better approximation for the unweighted, design and combined weighted methods. Regarding the regression estimates, the unweighted PC model gives values that are, on the average, within about 2% the Cox PH estimates, while the design and combined weighted methods give 3%.

As before, the variance estimates from the Cox PH model that are used for comparison are the robust variance estimates from the `coxph` and `cluster` options in R/SPlus, based on Lin and Wei [40] and do not include stratification. The unweighted variance estimation method gives values that are, on average, 0.6% close to the Cox PH based standard errors, the design cases give an average of 0.5%; and the combined method gives 1.8%.

The graphs in Figure 7.8 are analogous to the ones in Figure 7.5, discussed in the preceding section. They show the z-statistics $|\hat{\beta}|/se(\hat{\beta})$ calculated based on estimates from the Cox PH and PC models described above. The left and right graphs show estimates based on sampling design and on combined (design$\times$IPC) weights, respectively. As before, z-values based on robust variance estimates are labeled as "Cox PH.robust" and "PC.robutst", respectively. The label "PC.B&L" corresponds to design weighted variance estimates from the PC model, based on Boudreau and Lawless [9], the z-values labeled "PC.Comb" are based on the combined PC weighted variance estimates and those labeled "PC.Comb.Naive" are based on combined PC weighted variance estimates taking the IPC weights as fixed.

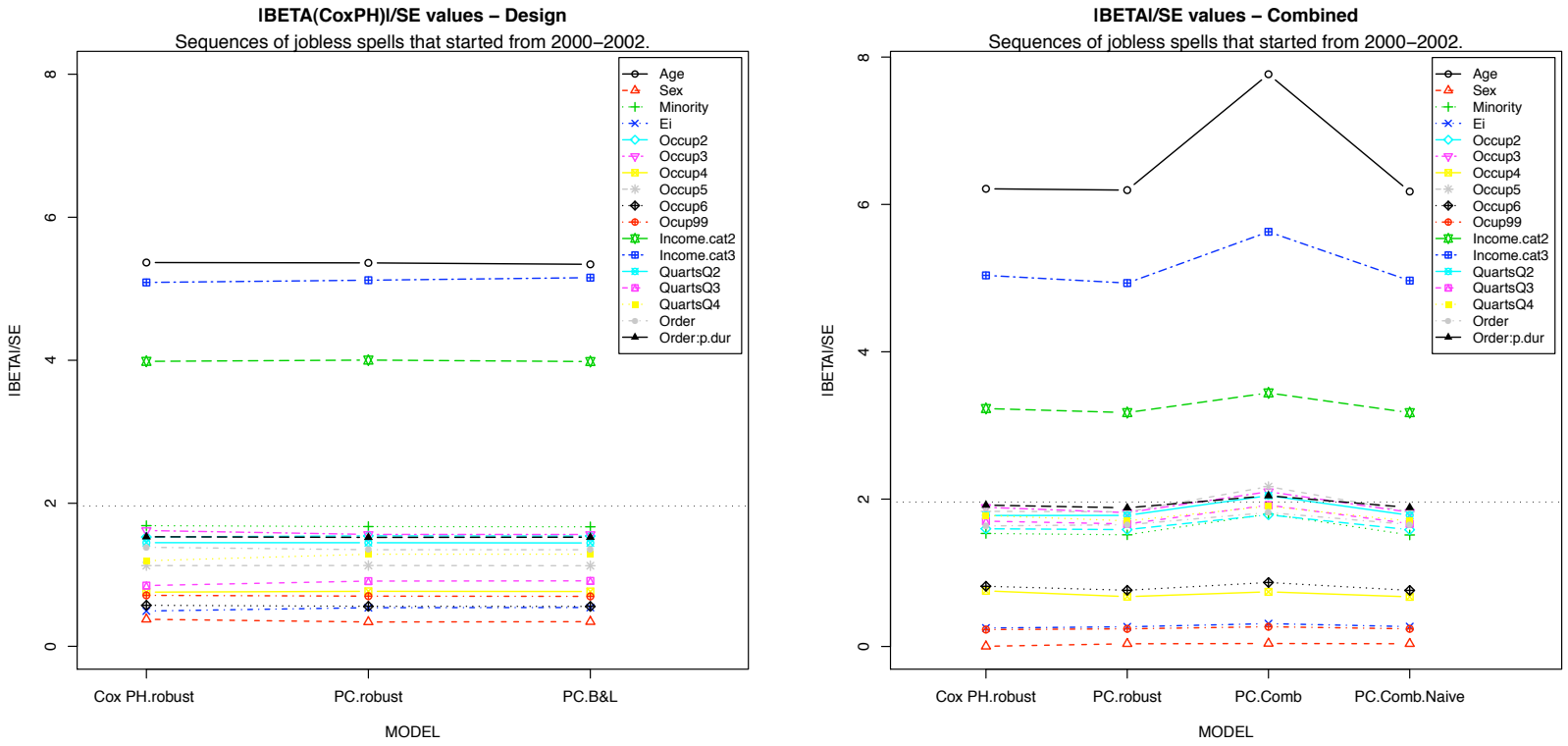

Figure 7.8: Values of $|\hat{\beta}|/se(\hat{\beta})$ from jobless spells starting in 2000-2002.

In terms of the significance of the variables, both design and combined methods lead to the same conclusions regarding Age and the two income categories Income.cat2 and Income.cat3, which are significant at a 5% level. The combined weighted methods, however, give greater z-values, especially for these three variables, and the significance of the Age variable is substantially higher than for the design methods. This increase of z-values implies a reduction of the variance estimates, which is consistent with what was observed in the preceding analysis. The z-values based on the robust variances give very similar results to those based on the PC.B&L in the design case and the PC.Comb.Naive methods in the combined case, implying that stratification is not important in these data. The combined methods place the variables Quarts, Minority, Order and the interaction Order:P.Dur closer to the 5% significance level than the design methods.

As a summary, the results shown in this section have allowed us to verify that the PC model gives a good approximation to the Cox PH model in real data sets for models with many variables. As expected, the PC.Comb method gives smaller variance estimates than PC.Comb.Naive, and in this case, the difference between these two does not lead to different conclusions regarding the significance of the variables. Another important conclusion is that the robust variance estimates without stratification (Cox PH.robust and PC.robust) are very similar to those based on PC.B&L and PC.Comb.Naive, which do account for strata.

Model selection and interpretation

A model with the variables in Table 7.13 including interactions between Age and Sex, Minority and Employment Insurance, Minority and Income, and Quarts and Income was assessed using the backwards elimination technique. From these interactions only Age:Sex remained significant in the model.

Table 7.17 shows the significance tests for the variables that were left in the final model using combined weights. The "Est." column refers to the regression estimates from the Cox PH model. The third and fourth columns refer to the standard errors and p-values based on the PC model ("PC.COMB"); similarly, the fifth and sixth columns are based on the robust variance estimate from the Cox PH model ("COX PH"). Results based on these two variance estimates show an overall agreement, both giving as most

significant the variables Age, Income and the term Order:P.Dur.

The Age coefficient is negative, agreeing with the modelling results on first jobless spells in 2000 shown in the preceding subsection. This indicates that as age increases, individuals are more likely to have longer jobless spells. The estimate for Age gives a hazard of $\exp(-0.28) = 0.97$ (95% CI of (0.96,0.98)), giving a 3% reduction in the hazard for leaving the jobless state for one year increment.

The Sex variable does not appear significant by itself, but it is significant at a 10% level when included in interaction with Age. Controlling for the other variables in the model, this interaction's estimated coefficient implies that as age increases in women, it becomes more likely for them to experience shorter spells than men.

With respect to the Minority variable, people that do not belong to a visible minority group have $\exp(0.311) = 1.36$ (95% CI of (0.99,1.88)) times the hazard for leaving the jobless state than those that do, that is, it is for them more likely to experience shorter spells, controlling for all other variables in the model.

The Occupation variable has a Wald test statistic of 30.82, which is is highly significant under the PC.COMB method. The individual p-values of the occupation types 3 (science, education, art) and 5 (processing, manufacturing) are significant at a 10% level, implying that the lengths of jobless spells from individuals in these occupation categories are significantly shorter than those that work in the reference category, given by type 1 (trades, transport, equipment operators). The hazard for individuals in level 5 have a $\exp 0.461 = 1.6$ relative risk of having shorter spells than people from level 1, with a 95% CI of (0.92, 2.72).

The variable income is highly significant, and the positive estimates imply shorter spells for those with highest income. In particular, subjects in the Income.cat3 category have a $\exp(0.93) = 2.53$ relative risk of experiencing shorter spells (95% CI of (1.84,3.50)) compared to the lowest income category.

Longer spells generally occur in a lower frequency than shorter spells. Significance of the Order variable would indicate that if the spell is of higher order then it is more likely to be shorter than if it is a first spell. This variable has a significance level close to the 10%. The interaction Order:P.Dur is significant at a 5% level, and its negative value

implies that among spells of higher order, there is a tendency of length increment as the preceding spell becomes longer.

Table 7.17: Summary from fitting unstratified Cox PH model to sequences of jobless spells in 2000-2002 from residents of Ontario in 1999.

		PC.COMB		COX PH.Robust (Naive)	
Variable	Est.	SE	p-Val	SE	p-Val
Age	-0.028	0.006	**<0.001**	0.007	**<0.001**
Sex	-0.064	0.126	0.611	0.134	0.630
Minority	0.311	0.162	**0.055**	0.188	**0.097**
Occup2	0.363	0.293	0.216	0.321	0.260
Occup3	0.462	0.276	**0.095**	0.311	0.140
Occup4	-0.416	0.356	0.242	0.382	0.280
Occup5	0.461	0.276	**0.095**	0.317	0.150
Occup6	0.137	0.292	0.638	0.325	0.670
Occup99	-0.248	0.331	0.455	0.366	0.500
Income.cat2	0.507	0.138	**0.000**	0.150	**0.001**
Income.cat3	0.930	0.164	**0.000**	0.185	**<0.001**
Order	0.136	0.100	0.173	0.112	0.230
Order:P.Dur	-0.010	0.005	**0.029**	0.005	**0.040**
Age:Sex	-0.018	0.010	**0.080**	0.011	0.110

Significant at a 10% level or less shown in bold.

Model checks

Informal model checks were done using the R function `cox.zph` as in the preceding section, and results in Table 7.18 show that many variables have a violation of this assumption, with a global test showing a highly significant departure with a p-value <0.0001.

Schoenfeld residual plots vs. time (or a function of time) should show a constant trend with no slope indicating that the coefficient does not depend on time, if the PH assumption is valid. Plots given by `cox.zph` were examined for all the variables (omitted for confidentiality), and a downward trend after approximately 52 weeks was observed in the variables with a 10% significant negative "rho", especially the Sex variable (Table 7.18).

Therneau and Grambsch [57] give a discussion about what can be done in the presence of PH model departures, and based on this and the observed trend, a second test was

performed on the set of spells, this time censoring those that were longer than 52 weeks. Results on the fit and proportional hazards assumption are shown in Table 7.19. The standard errors shown in this table correspond to the robust variance estimate obtained from the `coxph` output and the PH assessment is done based on them. These tests give indication that the model has improved with respect to the proportionality assumption.

Looking back at Kaplan-Meier estimate on Figure 7.7, it can be seen that the probability of a spell to have a duration longer than 52 weeks is of about 26%. Fom the results in Table 7.19, it is reasonable to assume that spells longer than 52 weeks have different characteristics, which should be taken into account in the model as an additional covariate or modelled separately.

An additional comment regarding the assessment of the model, is that the assumption of linearity of the Age variable was tested by adding a squared term, which did not result significant, and therefore evidence is insufficient to conclude the Age variable has a quadratic effect with these data.

Table 7.18: PH assessment, unstratified Cox PH fit on sequences of jobless spells in 2000-2002, ON.

Variable	rho	χ^2	p-Val
Age	-0.085	6.307	**0.012**
Sex	-0.118	18.071	**<0.001**
Minority	0.045	1.652	0.199
Occup2	0.023	1.005	0.316
Occup3	0.042	3.298	**0.069**
Occup4	0.046	2.901	**0.089**
Occup5	0.060	6.782	**0.009**
Occup6	0.006	0.072	0.789
Occup99	0.010	0.166	0.684
Income.cat2	-0.070	4.857	**0.028**
Income.cat3	0.090	9.288	**0.002**
Order	-0.052	2.357	0.125
Order:P.Dur	-0.084	8.550	**0.003**
Age:Sex	-0.076	7.951	**0.005**
GLOBAL	NA	69.857	**<0.0001**

Variables with significance of 10% or less are shown in bold.

Table 7.19: Summary of unstratified fit after censoring spells longer than 52 weeks and PH assessment, sequences of jobless spells in 2000-2002.

	Significance			PH Assessment		
Variable	Est.	SE	p-Val	rho	χ^2	p-Val
Age	-0.016	0.008	**0.064**	-0.031	0.936	0.333
Sex	0.060	0.142	0.670	-0.054	2.767	**0.096**
Minority	0.569	0.248	**0.022**	0.042	1.659	0.198
Occup2	0.175	0.301	0.560	-0.015	0.295	0.587
Occup3	0.374	0.287	0.190	0.009	0.102	0.749
Occup4	-0.350	0.326	0.280	0.013	0.163	0.687
Occup5	0.215	0.299	0.470	0.018	0.383	0.536
Occup6	0.177	0.297	0.550	-0.017	0.368	0.544
Occup99	-0.622	0.412	0.130	-0.004	0.022	0.882
Income.cat2	0.681	0.164	**<0.001**	-0.037	0.984	0.321
Income.cat3	0.987	0.206	**<0.001**	0.032	0.905	0.341
Order	0.216	0.132	0.100	-0.021	0.416	0.519
Order:P.Dur	-0.005	0.005	0.270	-0.046	1.912	0.167
Age:Sex	-0.018	0.012	0.130	-0.030	1.030	0.310
(GLOBAL)	-	-	-	NA	18.641	0.179

Variables significant at a 10% level or less are shown in bold.

Stratification

Cox PH stratified estimates will be compared with the unstratified estimates given in Tables 7.17 and Table 7.19. Recall that we are stratifying on the 11 strata given by economic regions and also considering these on the variance estimation, including the 389 clusters given by the dissemination areas (PSU's).

The variance estimate given in Boudreau and Lawless [9] is produced by the R/SPlus function `coxph` with the `strata` and `cluster` options and is used in the following analysis. Even though this variance estimate does not account for the random nature of the IPC weights, it can be used bearing in mind that the COMB Naive estimates may be higher than the COMB estimates.

The upper panel of Table 7.20 shows the results from fitting a stratified Cox PH model to the sequences of jobless spells in 2000-2002, together with the tests for the proportionality assumption. Looking back at Table 7.17 for the unstratified model, it can be seen that the effects of all variables remain significant at a 10% level, except for that of Minority which has a significant change in magnitude. The proportionality tests

in this table compared to those from the unstratified estimates (Table 7.18) show an improvement regarding the PH assumption.

Visual assessment of residual plots vs. time led us to assume that spells longer than 52 weeks could be causing these PH departures. The lower panel of Table 7.20 shows the results obtained after censoring spells that had durations longer 52 weeks. The most important variables in this fit are Age, Occup3, Occup4, Income, Order and Order:P.Dur. This is a better version of the unstratified model with estimates shown in Table 7.19. It is more informative in the sense that it has more significant variables and validates better the PH assumption for all variables.

Modelling by spell order

Kovacevic and Roberts [29] implement the Cox PH model in various forms to jobless spells from SLID, involving individuals in the panel from 1993 to 1998 across Canada. They analyze the first four spells in this period, and one of their analyses consists of modelling the spells separately by spell order, which is equivalent to doing a stratified Cox PH analysis, with different slopes and baseline hazards. They do not account for the dependence between spells within individuals explicitly in the model, but they account for it in variance estimation, using a robust estimate based on Lin [37], and compare results while using design based variance estimates based on Binder [4]. They find no substantial difference between these two variance estimation procedures in their particular example. Further, they do not account for dependent censoring.

In the discussion that follows, we model first and second jobless spells that started from 2000 to 2002 separately, from people residing in Ontario in 1999. Main effects models were fitted, and the variables used for first spells are age, sex, minority, employment insurance, occupation, income, yearly quarters and start year of the spell. The model for second spells further includes the duration of the preceding spell. The variables occupation, employment insurance and income from the year prior to the start of the spell with exception of the start year of the spell are described in Table 7.13, section 7.3.3.

There were 392 first and 136 second spells, from which 182 and 58 were censored, respectively. A model for third spells would have included 53 spells, but these did not

Table 7.20: Summary of stratified fit and PH Assessment, sequences of jobless spells in 2000-2002.

	Significance			PH Assessment		
Variable	Est.	SE	p-Val	rho	χ^2	p-Val
Age	-0.0265	0.0070	**0.0002**	-0.070	4.162	**0.041**
Sex	-0.0412	0.1352	0.7600	-0.062	4.580	**0.032**
Minority	0.1953	0.1912	0.3100	-0.008	0.052	0.820
Occup2	0.4522	0.3138	0.1500	0.016	0.484	0.487
Occup3	0.5352	0.3067	**0.0810**	0.036	2.349	0.125
Occup4	-0.5709	0.3558	0.1100	0.030	1.041	0.308
Occup5	0.5248	0.3133	**0.0940**	0.044	3.377	**0.066**
Occup6	0.1645	0.3214	0.6100	-0.003	0.022	0.882
Occup99	-0.2260	0.3428	0.5100	0.027	1.216	0.270
Income.cat2	0.5635	0.1519	**0.0002**	-0.014	0.198	0.657
Income.cat3	1.0104	0.1913	**<0.001**	0.069	5.623	**0.018**
Order	0.1134	0.1131	0.3200	-0.002	0.004	0.949
Order:P.Dur	-0.0089	0.0053	**0.0930**	-0.040	1.902	0.168
Age:Sex	-0.0185	0.0112	**0.0970**	-0.051	3.509	**0.061**
(GLOBAL)	-	-	-	NA	32.675	**0.003**

Summary after censoring spells longer than 52 weeeks

	Significance			PH Assessment		
Variable	Est.	SE	p-Val	rho	χ^2	p-Val
Age	-0.0189	0.0074	**0.0110**	-0.077	4.163	**0.041**
Sex	-0.0866	0.1275	0.5000	-0.059	2.570	0.109
Minority	0.2201	0.2178	0.3100	-0.047	1.698	0.193
Occup2	0.2958	0.2860	0.3000	0.011	0.146	0.702
Occup3	0.4390	0.2647	**0.0970**	0.020	0.374	0.541
Occup4	-0.5431	0.3091	**0.0790**	0.043	1.056	0.304
Occup5	0.1777	0.2655	0.5000	0.011	0.111	0.739
Occup6	0.1357	0.2503	0.5900	-0.032	0.960	0.327
Occup99	-0.1403	0.3318	0.6700	0.040	1.311	0.252
Income.cat2	0.8151	0.1640	**<0.0001**	-0.030	0.658	0.417
Income.cat3	1.0644	0.2074	**<0.0001**	-0.014	0.180	0.672
Order	0.2229	0.1238	**0.0720**	0.027	0.518	0.472
Order:P.Dur	-0.0080	0.0046	**0.0780**	-0.062	2.827	**0.093**
Age:Sex	-0.0057	0.0102	0.5700	0.041	1.263	0.261
(GLOBAL)	-	-	-	NA	18.415	0.189

Variables significant at a 10% level or less are shown in bold.

allow for stratification, so was left out of the analysis. There were less than 15 spells of the fourth and fifth order. Eleven strata and 389 clusters were used for estimation, as before, based on the economic regions and dissemination areas from SLID.

Tables 7.21 and 7.22 shows the results from fitting unstratified and stratified models to first and second spells, respectively. For first spells, significance of the variables of the unstratified model compared to the stratified model does not change dramatically; however, the PH test assumption gives better results for the stratified model. In the case of the second spells, the stratified model gives substantially different results than the unstratified, and this is also reflected in the PH assessment.

The primary difference between the models used by Kovacevic and Roberts and our models in Tables 7.21 and 7.22 , is that the latter account for dependent LTF. Keeping this in mind, and the fact that their models control for a different set of covariates (also, theirs do not take into account information about the spells, such as start time and previous spell duration), there can be found some similarities in the overall results. For instance, the variables income and age were both significant in most of their and our models. In our analyses, the Occupation variable does not have a high significance, which is also the case in Kovacevic and Roberts.

Modelling first and second spells separately, however, has its downside in that the employment experience across individuals most likely has started at different times. That is, second spells from separate individuals since the year 2000 may not be their second spells since the start of their employment processes, which may very likely, differ. As discussed in section 2.3, modelling by spell order in this context may have some descriptive value but is not ideal for analytic purposes.

Table 7.21: Summary of unstratified fit by jobless spell order and PH Assessment, sequences of jobless spells in 2000-2002.

First Spells

	Significance			PH Assessment		
Variable	Est.	SE	p-Val	rho	χ^2	p-Val
Age	-0.042	0.007	**<0.001**	-0.143	11.700	**0.001**
Sex	-0.053	0.143	0.710	-0.136	10.100	**0.002**
Minority	0.111	0.195	0.570	0.018	0.138	0.710
Ei	-0.018	0.232	0.940	-0.055	2.070	0.150
Occup2	0.424	0.236	**0.072**	-0.024	0.251	0.616
Occup3	0.115	0.262	0.660	-0.055	1.490	0.222
Occup4	-1.094	0.852	0.200	-0.018	0.286	0.593
Occup5	0.197	0.265	0.460	-0.021	0.195	0.659
Occup6	-0.085	0.250	0.730	-0.038	0.736	0.391
Occup99	-0.617	0.317	**0.052**	-0.072	2.480	0.115
Income.cat2	0.461	0.205	**0.024**	0.010	0.055	0.814
Income.cat3	0.866	0.223	**0.000**	0.103	4.830	**0.028**
Quarts2	-0.323	0.169	**0.057**	-0.091	4.260	**0.039**
Quarts3	-0.088	0.159	0.580	0.004	0.007	0.933
Quarts4	-0.493	0.191	**0.010**	-0.108	7.150	**0.008**
Stryr.cat3	-0.340	0.152	**0.025**	0.001	0.000	0.983
Stryr.cat4	-0.435	0.197	**0.027**	-0.058	2.150	0.143
(GLOBAL)	-	-	-	NA	42.000	**0.001**

Second Spells

	Significance			PH Assessment		
Variable	Est.	SE	p-Val	rho	χ^2	p-Val
Age	-0.027	0.012	**0.022**	-0.243	15.471	**0.000**
Sex	0.197	0.265	0.460	0.073	1.573	0.210
Minority	-0.183	0.311	0.560	-0.264	15.239	**<0.001**
Ei	-0.542	0.348	0.120	0.165	10.230	**0.001**
Occup2	0.479	0.434	0.270	0.051	0.764	0.382
Occup3	0.672	0.424	0.110	0.240	17.695	**<0.001**
Occup4	0.039	0.621	0.950	0.154	7.304	**0.007**
Occup5	0.424	0.484	0.380	0.128	5.237	**0.022**
Occup6	0.186	0.521	0.720	-0.092	3.167	**0.075**
Occup99	0.627	0.608	0.300	0.069	1.487	0.223
Income.cat2	1.058	0.311	**0.001**	-0.324	22.242	**<0.001**
Income.cat3	1.094	0.364	**0.003**	-0.098	2.377	0.123
Quarts2	-0.354	0.483	0.460	0.175	10.666	**0.001**
Quarts3	0.154	0.341	0.650	0.066	1.141	0.285
Quarts4	-0.059	0.407	0.880	0.232	18.864	**<0.001**
Stryr.cat3	0.122	0.268	0.650	0.261	14.119	**<0.001**
Stryr.cat4	0.455	0.344	0.190	0.128	4.859	**0.028**
P.Dur	-0.013	0.007	**0.042**	-0.199	9.899	**0.002**
(GLOBAL)	-	-	-	NA	92.458	**<0.001**

Variables significant at a 10% level or less are shown in bold.

Table 7.22: Summary of stratified fit by jobless spell order and PH Assessment, sequences of jobless spells in 2000-2002.

First Spells

	Significance			PH Assessment		
Variable	Est.	SE	p-Val	rho	χ^2	p-Val
Age	-0.042	0.007	**<0.001**	-0.106	6.300	**0.012**
Sex	-0.045	0.146	0.760	-0.056	1.750	0.186
Minority	0.124	0.211	0.560	0.025	0.265	0.607
Ei	-0.153	0.248	0.540	-0.086	5.230	**0.022**
Occup2	0.433	0.254	**0.088**	-0.006	0.020	0.887
Occup3	0.106	0.269	0.690	-0.043	0.897	0.344
Occup4	-0.976	0.756	0.200	-0.019	0.261	0.609
Occup5	0.160	0.260	0.540	-0.023	0.219	0.640
Occup6	-0.109	0.258	0.670	-0.022	0.261	0.609
Occup99	-0.631	0.308	**0.040**	-0.047	1.110	0.292
Income.cat2	0.735	0.206	**<0.001**	0.088	4.260	**0.039**
Income.cat3	1.122	0.249	**<0.001**	0.091	4.790	**0.029**
Quarts2	-0.380	0.186	**0.041**	-0.081	3.440	**0.064**
Quarts3	-0.101	0.167	0.550	0.000	0.000	0.998
Quarts4	-0.499	0.219	**0.023**	-0.059	2.560	0.110
Stryr.cat3	-0.387	0.153	**0.012**	-0.037	0.712	0.399
Stryr.cat4	-0.519	0.205	**0.011**	-0.081	4.450	**0.035**
(GLOBAL)	-	-	-	NA	25.200	**0.090**

Second Spells

	Significance			PH Assessment		
Variable	Est.	SE	p-Val	rho	χ^2	p-Val
Age	-0.024	0.014	**0.082**	-0.218	20.987	**<0.001**
Sex	0.416	0.247	**0.093**	0.102	2.545	0.111
Minority	-0.645	0.477	0.180	-0.262	31.415	**<0.001**
Ei	-0.575	0.388	0.140	0.159	10.790	**0.001**
Occup2	0.997	0.563	**0.076**	0.129	8.949	**0.003**
Occup3	1.319	0.621	**0.034**	0.139	11.495	**0.001**
Occup4	0.069	0.722	0.920	0.116	6.933	**0.008**
Occup5	0.865	0.605	0.150	0.123	8.376	**0.004**
Occup6	0.663	0.607	0.280	-0.039	0.853	0.356
Occup99	0.762	0.722	0.290	0.070	2.313	0.128
Income.cat2	0.897	0.326	**0.006**	-0.237	13.297	**<0.001**
Income.cat3	0.876	0.407	**0.031**	0.038	0.486	0.486
Quarts2	-0.640	0.641	0.320	0.117	6.028	**0.014**
Quarts3	0.012	0.426	0.980	0.120	4.447	**0.035**
Quarts4	-0.245	0.441	0.580	0.205	12.843	**<0.001**
Stryr.cat3	0.024	0.288	0.930	0.129	5.655	**0.017**
Stryr.cat4	0.351	0.373	0.350	-0.021	0.194	0.660
P.Dur	-0.014	0.006	**0.024**	-0.084	1.581	0.209
(GLOBAL)	-	-	-	NA	75.333	**<0.001**

Variables significant at a 10% level or less are shown in bold.

Chapter 8

Topics for Research

As a brief summary, the methods for the analysis of durations from longitudinal survey data developed in this thesis take into account for dependent loss to follow-up as a form of missing at random (MAR) mechanism. It has been shown that our methods provide unbiased estimates and that our variance estimates give better results when considering the IPC weights as random rather than fixed. The estimation techniques are based on theory for parametric methods, and their applicability in this thesis includes the estimation of regression coefficients from parametric survival models, the estimation of non-parametric survival distributions via the Kaplan-Meier estimate, and the estimation of regression coefficients and baseline cumulative functions from the Cox PH model through the piece-wise constant approximation model.

As mentioned in chapter 7, through the implementation of the methodology in jobless spells from SLID, we have encountered challenges regarding missing data and measurement error. In this chapter we give a short discussion on patterns of missing data and provide pointers towards dealing with them. We also give a discussion on missing data in response variables and covariates. Other topics that have been identified as areas for future research are discussed, such as generalizations of the MAR assumption, the development of methods to perform model checks and for the analysis of more general event history data.

The MAR assumption and missing data patterns

Violations of the missing at random (MAR) requirement for our methods are likely to occur in any survey in which panel members are seen at widely spaced interviews. Furthermore, as noted by Robins et al. [51], if we wanted to generalize our methods from a monotone missing data pattern to an intermittent one, then the data is no longer MAR. That is, $Pr(R_{it}|R_{i,t-1} = 1, H_i(M)) = Pr(R_{it}|R_{i,t-1} = 1, H_i(t-1))$ does not longer hold and we say that the data is not missing at random (NMAR). In this case it is necessary to make assumptions that cannot be checked in practice (Fitzmaurice et al. [19], part V), so auxiliary data from administrative sources or from tracing individuals who were lost to follow-up can be used. Another possibility is to perform sensitivity analysis to assess the effect of NMAR loss to follow-up or other types of missing data (e.g. Rotnitzky, Robins and Scharfstein [52], Scharfstein and Robins [56]).

Robins et al. [51], who applied IPCW methods in generalized linear models for longitudinal data in the non-survey context, provides a discussion on dealing with arbitrary patterns of missing information. Some other examples are Yi and Thompson [60], where a likelihood-based approach for longitudinal incomplete binary data with possibly NMAR drop-outs is discussed. Yi and Cook [59] discuss an extension to deal with intermittently missing data, and present an application to a longitudinal cluster-randomized smoking prevention trial. The development of methods for NMAR assumption and intermittent missing data patterns is a needed area of study in the context of duration analysis from longitudinal survey data.

Missing data in response variables and covariates

In the analysis of jobless spells from SLID, we have encountered a variety of types of missing data besides the one caused by attrition, which were dealt with, but methods for other types of missing data need to be proposed. The introduction of chapter 7 gives a detailed discussion on missing data from SLID, and as mentioned, one type has to do with the start dates of the spells. Table B.4 in the appendix, shows that about 41% of the individuals who had at least one jobless spell in 1999-2004 had at least one spell with an unknown start date. In the data set that was used for analysis of jobless spells, there were about 26% of spells with an unknown start date, as shown in Table 7.5. Discarding

the spells that had an unknown start date may severely bias the estimation results and it has further impact in the modelling of dropout, through covariates that are based on information like the length of a preceding spell.

Another type of missing information has been found in variables that are not directly related to the start and end of jobless spells, but are intended to be used as covariates for duration or dropout models. In SLID, this missing information comes in the form of responses of the types "don't know", "not available" or "refusal". As a way to deal with this, even though it is not ideal, missing values for categorical covariates were included in most models as an additional covariate level.

A comprehensive discussion of methods for dealing with missing data can be found in Little and Rubin [43]. A technique known as the complete case method consists of simply using only individuals with complete data. This strategy may be satisfactory with small amounts of missing data and when data are missing completely at random (MCAR), but can lead to serious bias under MAR or NMAR conditions. Weighting procedures, which can be seen as a form of the complete case method, such as those used in SLID that account for non-response, consist of adjusting the sampling design weights and then analyzing the complete units. Imputation methods consist of filling in the missing values and then the imputed and observed data are analyzed together as "complete" data. Single imputation has the disadvantage that imputing a single value treats that value as known, and without special adjustments single imputation cannot reflect sampling variability. Multiple imputation (Rubin [54]) is a method that incorporates imputation uncertainty and is preferred over simple imputation when there are large amounts of missing data. The estimates of variance based on multiple imputation methods are based on a specific model and consider a particular missing data mechanism. In our case for example, a model in which the distribution of the start time of a spell depends on information from individuals and also on end times of the spells could be implemented and then values may be randomly imputed, based on this model, to spells that have a missing start date. Challenges of multiple imputation and approaches towards validating assumptions are discussed in Kenward and Carpenter [27].

There are maximum likelihood and Bayes methods available when a model for all variables (e.g. covariates, truncation times, start times, durations) is available and data

are MAR. Cook and Lawless [15] (section 8.6) give a discussion on maximum likelihood methods for missing covariates along with references. Regarding the Cox model in survival analysis, Kalbfleisch and Prentice [25] (section 11.5) give some references. It is possible to extend methods in Robins et al. [51], Lawless et al. [34] and related references to deal with survey data but to date this has not been done for settings involving duration analysis.

Measurement error

One example of measurement error, in SLID and in many longitudinal surveys, is the seam effect. It is a form of recall bias and refers to a high occurrence of reported transitions at the seam between two reference periods, in our case the SLID waves. The SLID labour interviews are performed in January of each year and information is collected about the individuals labor activity during the preceding year, called the reference year. The interviewed person may have a better recollection of his or her labor activity from the second half of the year than from the first half, for instance. This may induce a biased response towards the beginning of a reference year. It is important to consider this seam effect issue when drawing conclusions from the data. Discussion of the extent and consequences of the seam effect in longitudinal surveys can be found in Callegaro[10], and in particular for SLID in Cotton and Gilles [16] and Lemaitre [36].

Kalton et al. [26] provide a discussion regarding adjustments for seam effect while analyzing spells from the US Survey of Income and Program Participation (SIPP). They propose an adjustment that produces smooth distributions of starts and ends of the spells. They compute weighted Kaplan-Meier estimates and identify the time in which the seam effect takes place, and then reallocate a proportion of the reported starts and ends at the seam effect to other time periods within the duration of the wave. Kalton et al. discuss the cases of single and multiple spells per individual. It would be interesting to examine the seam effect on SLID data, and develop methods since contributions in dealing with this appear limited.

Another type of measurement error was found in variables regarding the termination of the jobless spells. There are spells in SLID that are not observed to completion, and to account for this, there is a SLID variable ("endtyp7") that provides information

associated with the end of a jobless spell, indicating whether the respondent reported working in subsequent interviews, there was non-response or the person was no longer eligible for the labour interview. We have associated the two latter possibilities to our dropout related variable for estimation of the IPC weights, but the question remains on what is the best way to deal with the former, especially when spells that end this way occurred within a sequence (see section B.4 in the appendix and Table B.6).

Model checks

Classical duration analysis theory has a number of model checking methods, however, this topic has not been developed for longitudinal survey data or when weights are treated as random. Nonetheless, diagnostic checks by treating the weights as fixed can still be useful. That is, to perform usual model checks by Cox PH software, like constructing plots of weighted martingale residuals or DFBETAS, and using tests for the proportionality assumption via `cox.zph` in R/Splus using naive variance estimates.

More complex processes

The methods developed here involve two alternating processes, illustrated by unemployment and employment durations. These methods can be easily extended to deal with more general types of event history analysis. For example, in the competing risks model (see section 1.3), an exit from a state can occur with a transition to one or more possible states, and the hazard functions related to the transitions can be treated as transition intensity functions (Andersen et al. [1]). For example, exits from the "unemployed" (UE) state can occur by transitioning to the "employed" (E) or to the "out of the labour force" (O) states. In the competing risks framework, different types of transitions can be modeled separately (Lawless [32], ch. 9), so when analyzing transitions from the UE to the E state, the transitions from UE to O can be treated as a censoring event and the methods developed here can be readily applied. Processes where individuals can make transitions back an forth between states are more complex and constitute an opportunity for research.

Appendix A

Simulation Results for Kaplan-Meier Estimation

Table A.1: Empirical survival probabilities in population.

Scenario I		Scenario II		Scenario III		Scenario IV	
Time	Surv.	Time	Surv.	Time	Surv.	Time	Surv.
9	0.9144	10	0.9044	9	0.9234	9	0.9198
12	0.8232	13	0.8174	13	0.8082	12	0.8292
15	0.7239	16	0.7165	16	0.7063	15	0.7318
18	0.6202	19	0.6168	19	0.6084	18	0.6292
21	0.5231	22	0.5241	22	0.5166	22	0.5028
25	0.4115	26	0.4152	26	0.4089	25	0.4211
30	0.3019	31	0.3095	31	0.3028	30	0.3112
36	0.2087	38	0.2019	37	0.2115	37	0.2013
47	0.1063	49	0.1036	49	0.1040	48	0.1044

Table A.2: Bias and Empirical Standard Error for $\hat{S}(y)$, scenario I.

	Bias		Empirical SE	
Time	DES	COMB	DES	COMB
9	-0.0116	0.0004	0.0152	0.0137
12	-0.0200	0.0006	0.0198	0.0187
15	-0.0253	0.0016	0.0230	0.0223
18	-0.0305	0.0014	0.0238	0.0239
21	-0.0322	0.0015	0.0238	0.0248
25	-0.0313	0.0021	0.0234	0.0254
30	-0.0292	0.0016	0.0210	0.0239
36	-0.0246	0.0013	0.0182	0.0218
47	-0.0165	0.0010	0.0134	0.0174

Table A.3: Average standard error and coverage for $\hat{S}(y)$, scenario I

	Average SE				Coverage			
	DES		COMB		DES		COMB	
Time	Sup.	Fin.	Comb.	Naive	Sup.	Fin.	Comb.	Naive
9	0.0143	0.0141	0.0126	0.0129	0.8180	0.7910	0.9300	0.9450
12	0.0190	0.0187	0.0175	0.0178	0.7880	0.7670	0.9250	0.9370
15	0.0218	0.0215	0.0207	0.0212	0.7550	0.7460	0.9320	0.9370
18	0.0233	0.0229	0.0228	0.0233	0.7300	0.7080	0.9400	0.9450
21	0.0236	0.0231	0.0239	0.0244	0.7040	0.6990	0.9470	0.9590
25	0.0228	0.0223	0.0239	0.0244	0.7120	0.7020	0.9470	0.9550
30	0.0208	0.0205	0.0228	0.0233	0.7120	0.7160	0.9320	0.9380
36	0.0182	0.0178	0.0208	0.0212	0.7347	0.7447	0.9327	0.9427
47	0.0135	0.0133	0.0165	0.0169	0.7895	0.8206	0.9377	0.9398

Table A.4: Bias and Empirical Standard Error for $\hat{S}(y)$, scenario II.

	Bias		Empirical SE	
Time	DES	COMB	DES	COMB
10	0.00096	0.00123	0.01516	0.01587
13	0.00085	0.00145	0.01964	0.02041
16	0.00081	0.00160	0.02282	0.02361
19	0.00117	0.00236	0.02474	0.02573
22	0.00134	0.00227	0.02508	0.02634
26	0.00130	0.00227	0.02489	0.02597
31	0.00337	0.00366	0.02250	0.02360
38	0.00360	0.00384	0.01939	0.02018
49	0.00160	0.00148	0.01488	0.01569

Table A.5: Average standard error and coverage for $\hat{S}(y)$, scenario II

	Average SE				Coverage			
	DES		COMB		DES		COMB	
Time	Sup.	Fin.	Comb.	Naive	Sup.	Fin.	Comb.	Naive
10	0.0141	0.0140	0.0146	0.0147	0.9220	0.9310	0.9240	0.9380
13	0.0185	0.0183	0.0191	0.0192	0.9390	0.9340	0.9420	0.9400
16	0.0215	0.0212	0.0222	0.0223	0.9320	0.9260	0.9390	0.9360
19	0.0231	0.0228	0.0239	0.0240	0.9340	0.9350	0.9380	0.9400
22	0.0237	0.0234	0.0246	0.0246	0.9430	0.9410	0.9310	0.9400
26	0.0233	0.0230	0.0242	0.0243	0.9460	0.9420	0.9320	0.9390
31	0.0219	0.0216	0.0227	0.0228	0.9419	0.9389	0.9389	0.9419
38	0.0190	0.0188	0.0197	0.0198	0.9418	0.9388	0.9438	0.9438
49	0.0146	0.0144	0.0150	0.0151	0.9460	0.9384	0.9449	0.9438

Table A.6: Bias and Empirical Standard Error for $\hat{S}(y)$, scenario III.

	Bias		Empirical SE	
Time	DES	COMB	DES	COMB
9	-0.0029	0.00005	0.0136	0.0137
13	-0.0055	0.0008	0.0205	0.0205
16	-0.0063	0.0012	0.0235	0.0237
19	-0.0070	0.0016	0.0239	0.0245
22	-0.0066	0.0020	0.0235	0.0246
26	-0.0064	0.0017	0.0228	0.0244
31	-0.0059	0.0017	0.0210	0.0226
37	-0.0052	0.0011	0.0188	0.0206
49	-0.0029	0.0010	0.0138	0.0151

Table A.7: Average standard error and coverage for $\hat{S}(y)$, scenario III.

	DES		COMB		DES		COMB	
Time	Sup.	Fin.	Comb.	Naive	Sup.	Fin.	Comb.	Naive
9	0.0131	0.0129	0.0130	0.0130	0.9320	0.9290	0.9280	0.9460
13	0.0191	0.0188	0.0192	0.0192	0.9100	0.9070	0.9260	0.9300
16	0.0218	0.0216	0.0223	0.0224	0.9250	0.9170	0.9300	0.9370
19	0.0233	0.0230	0.0240	0.0240	0.9230	0.9220	0.9340	0.9470
22	0.0237	0.0233	0.0246	0.0246	0.9380	0.9380	0.9480	0.9510
26	0.0232	0.0228	0.0243	0.0243	0.9380	0.9410	0.9450	0.9450
31	0.0215	0.0212	0.0228	0.0228	0.9490	0.9520	0.9580	0.9590
37	0.0190	0.0188	0.0204	0.0204	0.9458	0.9518	0.9518	0.9528
49	0.0142	0.0140	0.0154	0.0154	0.9511	0.9565	0.9489	0.9500

Table A.8: Bias and Empirical Standard Error for $\hat{S}(y)$, scenario IV.

	Bias		Empirical SE	
Time	DES	COMB	DES	COMB
9	-0.0078	0.00004	0.0144	0.0136
12	-0.0133	0.0007	0.0199	0.0191
15	-0.0181	0.0004	0.0226	0.0222
18	-0.0204	0.0007	0.0234	0.0238
22	-0.0210	0.0017	0.0235	0.0246
25	-0.0209	0.0014	0.0230	0.0244
30	-0.0192	0.0015	0.0211	0.0230
37	-0.0158	0.0012	0.0178	0.0204
48	-0.0099	0.0010	0.0134	0.0160

Table A.9: Average standard error and coverage for $\hat{S}(y)$, scenario IV

	Average SE				Coverage			
	DES		COMB		DES		COMB	
Time	Sup.	Fin.	Comb.	Naive	Sup.	Fin.	Comb.	Naive
9	0.0136	0.0135	0.0128	0.0128	0.8800	0.8670	0.9340	0.9420
12	0.0185	0.0183	0.0178	0.0179	0.8620	0.8390	0.9320	0.9360
15	0.0215	0.0212	0.0212	0.0213	0.8370	0.8210	0.9370	0.9430
18	0.0231	0.0228	0.0234	0.0234	0.8490	0.8370	0.9470	0.9520
22	0.0236	0.0232	0.0245	0.0245	0.8400	0.8390	0.9480	0.9500
25	0.0231	0.0227	0.0244	0.0245	0.8430	0.8440	0.9570	0.9580
30	0.0213	0.0210	0.0232	0.0233	0.8600	0.8630	0.9560	0.9560
37	0.0183	0.0180	0.0206	0.0206	0.8744	0.8854	0.9558	0.9548
48	0.0138	0.0136	0.0160	0.0160	0.8935	0.9204	0.9441	0.9548

Appendix B

Definitions and Exploratory Statistics from SLID

B.1 Defining loss to follow-up in SLID

LTF in exploratory discussion, section 7.2

Being lost to follow-up involves whether a person was in or out of scope and whether there was labour information available for that person. The term "in scope" in a particular year means that, as of December 31 of that year, the person was not deceased, lived in one of the ten Canadian provinces, did not live on an Indian reserve, had not been living in an institution for more than six months, or was not a full time member of the Canadian Armed Forces living in military barracks. There are four possibilities for missing labour information found in SLID data:

i. The person did not respond in the reference year but is still in scope. The person is considered a soft refusal since it might be possible to obtain data from them in a future year (SLID variables: ailgwt26=0, resp99=01))

ii. The person is in scope but no labour information is available in that year (SLID variables: resp99=01, nbjbs28=97 - "don't know")

iii. The person is out of scope (SLID variable: resp99=02-06)

iv. The person dropped out from the sample (SLID variable: resp99=07)

Over the six years from 1999 to 2004, individuals may have experienced one or more of the above possibilities (i)-(iii), while (iv) may have been experienced only once, since a person that drops from the sample is not included subsequently. An individual's labour history in the six years may have a combination of (i)-(iii), forming patterns that may be intermittent over the six years.

Persons that experienced any of the above conditions in the first two consecutive years of the panel (1999 and 2000) were excluded from the analysis. As a convention, everyone in the working data set was observed and had labour information in the first year of the panel and was followed until the first year in which any of the above was experienced. This year will be denoted as the **loss to follow-up (LTF) time**, or year. Individuals that had (i), (ii) or (iii) in the first year (1999) but that were observed and had labour information in the second year were applied the same follow-up definition citeria starting from 2000.

LTF in Kaplan-Meier estimation and Cox PH modelling, 7.3

The descriptive information presented in section 7.1 pertains to loss to follow up as defined above. A further restriction for LTF was used in sections 7.2 and 7.3, where Kaplan-Meier estimation and Cox PH modelling are performed. Individuals' time to LTF was further adjusted according to patterns of "looking for work" and "not looking for work" observed in the individual jobless spells sequences.

Table B.1 shows the number of spells in the working data set by "looking for work" response. There are about 35% of the spells in which the person was not looking for work. Jobless spells where the person was not looking for work have a different distribution than those from people who were looking. This was verified in a short analysis (not shown here) where distributions of jobless spells with "looking for work" and "don't know" were similar and stood apart from the distributions of spells with "not looking". Sequences of the kind "0 7 2" for example, were truncated to "0 7" and LTF was adjusted accordingly.

Table B.1: Number of jobless spells by "looking for work" response, after adjustments.

Spell order	1 (Yes)	2 (No)	7 (Don't know)	Total
1	4652	3629	3600	11881
2	2414	2017	297	4728
3	1166	913	133	2212
4	585	428	44	1057
5+	488	270	33	791
Total	9305	7257	4107	20669

B.2 Jobless spell SLID definition

The labour force status of a person at a given time can be assigned to one of the following categories:

a) Have a job and working.

b) Self-employed or unpaid family worker.

c) Have a job but absent for something other than a layoff or waiting for job to start.

d) Have a job but absent due to layoff or waiting for job to start.

e) Does not have a job but looking for work.

f) Remainder (do not have a job and not looking).

The SLID (and Labour Force Survey) definition of employed refers to sets (a), (b) and (c); unemployed refers to (d) and (e); and not in the labour force to (f) (not employed or unemployed).

A "jobless spell" in SLID, is defined as the period of time in which a person is out of work and may or may not be looking for work, categories (e) and (f). The descriptive analysis in section 7.1 pertains to this definition while the analyses in sections 7.2 and 7.3 pertains to the jobless spells where the person was looking for work (category (e)).

B.3 General features of SLID sample

Among the 31576 observed spells from the third SLID panel, about 73.4% had an known start date. From these spells, Table B.2 shows the number of spells by start year and number of spells that started before January of 1999.

Table B.2: Jobless spell counts by start year, SLID data panel 3.

Year	Frequency	Proportion	Cumulative Frequency	Cumulative Proportion
1999	3,333	0.1433	3,333	0.1433
2000	2,789	0.1199	6,122	0.2632
2001	3,096	0.1331	9,218	0.3964
2002	2,725	0.1172	11,943	0.5135
2003	2,662	0.1145	14,605	0.6280
2004	2,505	0.1077	17,110	0.7357
Before 1999	6,146	0.2643	23,256*	1.0000

* This amounts for 73.4% of spells with a known start date.

Table B.3 shows counts of individuals by number of jobless spells; 59.3% of the individuals had only one spell, 22.11% had two, and the remaining 18.56% consists of individuals having three jobless spells or more. The total number of individuals that had at least one jobless spell is 18,009. The number of individuals with no spells in the six-year period is $43683 - 18009 = 25674$, 58% from the total number of individuals.

Table B.4 shows counts of individuals by their number of spells with unknown starting dates. For instance, among the total of 3981 persons with two jobless spells (last column), there are 1764 persons for whom both spells have a known starting date. There were 1828 individuals with two spells, 1018 and 810 had the first and second spell with a missing start date, respectively. Further, there were 389 people with their two spells with an unknown start date.

The total number of individuals that had all their starting spells dates known is 10551, those that had an unknown start date once are 6635 and more than once are 823. These all sum up to the total number of people who had at least one jobless spell in the six

Table B.3: Number of individuals by number of spells, SLID data panel 3.

No. of spells	Frequency	Proportion	Cumulative Frequency	Cumulative Proportion
1	10,685	0.5933	10,685	0.5933
2	3,981	0.2211	14,666	0.8144
3	1,723	0.0957	16,389	0.9100
4	871	0.0484	17,260	0.9584
5	443	0.0246	17,703	0.9830
6	182	0.0101	17,885	0.9931
7	67	0.0037	17,952	0.9968
8	31	0.0017	17,983	0.9986
9+	26	0.0014	18,009	1.0000

years, 18009. Further, those individuals with all their starting jobless spell dates known (10551) have altogether a total of 16627 spells.

B.4 Ending types of jobless spells from SLID.

The definition of LTF from SLID variables from section B.1 can used to determine the year in which a person was last seen. That is, if a person was considered lost to follow-up for example, in 2001, then he or she was last seen the year before, in 2000. A jobless spell from this person that was ongoing in the year 2000 is therefore not observed completely at the end of this year, and is labeled as censored.

There is a SLID variable "endtyp7", that indicates whether the jobless spell was observed completely or not. That is, endtyp7=1 if a spell ended "normally", that is, was completely observed; and endtyp7=2 if ended "not normally", which may be due to any of the following reasons: (a) the respondent reported working in subsequent interviews, (b) there is non-response or (c) the respondent is no longer eligible for the labour interview.

The reason (a) refers to an inconsistency in the responses from individuals in two consecutive interviews, and happens when during the interview of a given year, the person reported that was jobless in a time period and the next year's interview reports that was

Table B.4: Number of individuals with known and unknown durations in their sequence of jobless spells.

No. spells	All within-individual spells with a known start date	Within-individual spells with unknown start dates						Total
		Once					More than once	
		1st	2nd	3rd	4th	5th$^+$		
1	7,281	3,404	-	-	-	-	-	10,685
2	1,764	1,018	810	-	-	-	* 389	3,981
3	800	486	124	112	-	-	** 201	1,723
4	375	212	43	63	46	-	132	871
5$^+$	331	179	27	27	45	39	101	749
Total	10,551	5,299	1,004	202	91	39	823	18,009

* Both spells with unknown start date.

** 20 individuals with three spells with unknown start date.

working in that period.

In general, jobless spells with endtyp7=2 end in the same year in which the person was last seen. However, this is not always the case. It is assumed that our LTF variable accounts for reasons (b) and (c) above, because every end date of spells that coincides with the year the person was last seen has endtyp7=2. Whenever the end date does not match with the year the person was last seen, it will be assumed that the reason for the spell being incomplete was (a). As a convention, every spell that has a type 2 ending will be treated as censored if its date coincides with the year last seen; but if the spell is succeeded by more spells then it will be treated as fully observed.

This is illustrated in Table B.5. Note that the second spell from person "X" has a type 2 ending (endtyp7=2: not normally) and is succeeded by another spell, so it is coded as "observed". The last spell ends the same year in which the person was last seen, therefore will be treated as censored. The only spell of person "Y" had a type 2 ending but it did not end the same year in which the person was last seen, so was coded as "observed".

This decision sounds reasonable from the perspective that if the spell was recorded as

Table B.5: Illustration of Status (censoring variable) depending on ending types of jobless spells and year prior to LTF.

Personid	Start date	End date	Spell ID	Endtyp7	Year last Seen	Status (0=Obs.)
X	05/09/1999	10/11/1999	1	1	2001	0
X	05/08/2000	31/12/2000	2	2	2001	0
X	23/11/2001	31/12/2001	3	2	2001	1
Y	10/11/2003	31/12/2003	1	2	2004	0

incomplete because the respondent reported working in subsequent inteviews, this can be considered as an interview collection mistake and thus may form part of the variability associated with this type of error.

There were 1759 spells that ended at least one year prior to the year in which the person was last seen. Table B.6 shows the number of complete and incomplete spells by the endtyp7 variable and by starting year.

Table B.6: Number of spells by starting year and ending type based on the SLID variable endtyp7. "EndYear1": spell end year = last year seen; "EndYear2": spell end year < last year seen.

START		ENDTYP7=2		
YEAR	ENDTYP7=1	EndYear1	EndYear2	TOTAL
1999	2,243	672	416	3,331
2000	1,928	463	398	2,789
2001	1,844	516	387	2,747
2002	1,574	436	316	2,326
2003	1,358	494	242	2,094
2004	669	1,239	0	1,908
Total	9,616	3,820	1,759	15,195

Appendix C

Modelling Loss to Follow-up from SLID

C.1 Variables used

The model that was used to describe dropout from SLID is the logistic model discussed in section 4.1, expression (4.5). The selection of covariates was based on those that are used for non-response modeling in SLID, see La Roche [30]. In SLID, non-response is one of the four conditions that define our dropout response variable, discussed previously. The list of covariates and counts of individuals within each category are shown in Tables C.1 and C.2 for Ontario and Quebec, respectively.

An additional level referring to missing values ("don't know", "refusals") is included in the variables Education Level, Immigration Status, Student and Jobless Status. Records with missing values in the variable Marital Status were not included, since they accounted for few people, about 15 in years 1999 and 2000 together. The number of individuals within each missing category are also shown in tables C.1 and C.2.

C.2 Summary of model fits

The model selection was performed using the backwards elimination technique, which is done automatically in SAS. The selection of variables is done based on Wald tests. Tables

Table C.1: Number of individuals by covariate category, LTF model (Ontario).

		Year				
Variable	Level	2	3	4	5	6
Sex	1(male)	2,321	2,303	2,006	1,682	1,560
	2(female)	2,091	2,129	1,874	1,577	1,464
Age *	Continuous centered	4,412	4,432	3,880	3,259	3,024
Edlev	L(low)	834	701	624	486	447
	LM(low-med)	603	533	492	454	431
	M (med)	2,738	2,368	2,140	1,857	1,748
	H (high)	185	152	143	124	116
	Missing	52	678	481	338	282
Marst **	1(married/comon law)	2,960	2,935	2,578	2,245	2,073
	2(single)	1,097	1,108	951	710	657
	3(other)	350	388	351	304	294
Immst	1(yes)	897	712	626	525	479
	2(no)	3,508	3,072	2,803	2,416	2,281
	Missing	<15	648	451	318	264
Stud	0 (not a student)	3,720	3,694	3,249	2,802	2,580
	1(full time)	492	514	444	306	320
	2(part time)	182	207	176	142	106
	Missing	18	17	<15	<15	18
Renter	1(yes)	3,461	3,553	3,186	2,749	2,597
	2(no)	951	879	694	510	427
	Other	0	25	35	29	33
HHsz	1	372	396	302	262	244
	2	1,068	1,051	913	772	722
	3	899	970	857	728	673
	4	1,323	1,224	1,120	935	858
	5+	750	791	688	562	527
Famtype	1(unrelated person)	483	485	373	306	283
	2(couple/lone no child)	820	834	734	639	583
	3(couple/lone child)	2,449	2,406	2,128	1,770	1,656
	4(other)	660	707	645	544	502
HHtype	1(one family)	4,264	4,305	3,767	3,179	2,957
	2(multi-family)	148	127	113	80	67
Urban	1(yes)	3,603	3,600	3,141	2,633	2,393
	2(no)	809	832	739	626	631
Jstat	1(jobless)	3,744	3,552	3,185	2,777	2,599
	2(not jobless)	395	470	357	274	239
	Missing	273	410	338	208	186
Total individuals		4,412	4,432	3,880	3,259	3,024

* Mean values of age by year: 40.17,40.93,41.46,42.54,42.98

** Less than 15 missing values in years 2 and 3 together.

Missing values are used as covariate category except for marst.

Table C.2: Number of individuals by covariate category, LTF model (Quebec).

		Year				
Variable	Level	2	3	4	5	6
Sex	1(male)	1,606	1,443	1,228	1,056	970
	2(female)	1,496	1,336	1,169	1,013	928
Age *	Continuous centered	3,102	2,779	2,397	2,069	1,898
Edlev	L(low)	833	672	558	458	424
	LM(low-med)	341	300	281	255	240
	M (med)	1,797	1,480	1,327	1,179	1,077
	H (high)	77	67	63	58	58
	Missing	54	260	168	119	99
Marst **	1(married/comon law)	1,976	1,770	1,529	1,333	1,188
	2(single)	841	725	595	463	436
	3(other)	284	284	273	273	274
Immst	1(yes)	161	123	103	91	75
	2(no)	2,936	2,429	2,157	1,881	1,744
	Missing	<15	227	137	97	79
Stud	0 (not a student)	132	111	113	89	59
	1(full time)	353	321	264	196	184
	2(part time)	2,604	2,327	2,004	1,783	1,648
	Missing	<15	20	16	<15	<15
Renter	1(yes)	2,306	2,050	1,787	1,559	1,477
	2(no)	796	729	610	510	421
HHsz	1	357	314	254	243	227
	2	779	762	663	596	558
	3	722	626	534	437	386
	4	821	718	636	550	491
	5+	423	359	310	243	236
Famtype	1(unrelated person)	420	370	294	278	262
	2(couple/lone no child)	542	539	482	436	409
	3(couple/lone child)	1,806	1,579	1,349	1,120	1,014
	4(other)	334	291	272	235	213
Hhtype	1(one family)	3,022	2,711	2,345	2,027	1,853
	2(multi-family)	80	68	52	42	45
Urban	1(yes)	2,236	1,979	1,725	1,483	1,334
	2(no)	866	800	672	586	564
Jstat	1(jobless)	2,465	2,119	1,861	1,697	1,564
	2(not jobless)	378	351	290	199	176
	Missing	259	309	246	173	158
Total individuals		3,102	2,779	2,397	2,069	1,898

* Mean values of age by year: 40.44,41.49,42.17,43.28,43.78

Missing values are used as covariate category.

C.3 to C.5 and C.6 to C.8 show the summary of the fits for years 2 to 6 (2000 to 2004) based on the data sets from Ontario and Quebec, respectively.

C.3 Model checks.

The assessment of the model was performed in SAS, and is based on the method by Hosmer and Lemeshow [24] for ungrouped binary data, by constructing levels of risk calculated from the estimated logistic model and comparing observed versus expected frequencies within each level. The levels of risk were obtained by sorting the fitted values from the dropout model in ascending order and then dividing them into 10 groups. These percentile groups are formed so that each contains approximately one tenth of the data and are known as "deciles of risk".

Within each group and for each value of $t \in \{2, ..., 6\}$, the expected and observed number of responses where $R_t = 1$ and $R_t = 0$ were compared. The members of the first group correspond to the lowest estimated probabilities that $R_t = 1$ or $R_t = 0$, and so on.

The expected number of events where $R_t = 1$ is estimated by $\hat{e}_k = n_k \bar{p}_k$, where $\bar{p}_k = \sum_{j=1}^{n_k} \hat{p}_j$ is the mean logistic probability for each group and $n_k \approx n/10$, where n represents the number of observations. Similarly, the expected and observed number of events where $R_t = 0$ are $n_k - \hat{e}_k$ and $n_k - o_k$, respectively.

Tables C.9 to C.10 and C.11 to C.12 show output from the assessment of the LTF models for $t \in \{2, 3, ...6\}$, for Ontario and Quebec, respectively. From the five tables, it can be seen that the expected number of observations within each group remains fairly close to the observed. Tables C.13 and C.14 confirm this, showing high p-values based on the Hosmer and Lemeshow statistic by year and province.

Table C.3: Summary of model fits, LTF, years 2 and 3 (Ontario).

Year	Variable	Level	DF	Estimate	StdErr	WaldChiSq	ProbChiSq
2	Intercept		1	0.6975	0.1804	14.9543	0.0001
2	Age		1	-0.0032	0.0062	0.2661	0.6060
2	Age^2		1	-0.0008	0.0003	5.8144	0.0159
2	Edlev	2:LM	1	0.3231	0.1558	4.2994	0.0381
2	Edlev	3:M	1	0.2419	0.1123	4.6430	0.0312
2	Edlev	4:H	1	0.0659	0.2184	0.0911	0.7628
2	Edlev	5:97	1	-0.6939	0.3434	4.0836	0.0433
2	Marst	2	1	-0.4289	0.1259	11.6052	0.0007
2	Marst	3	1	-0.4719	0.1703	7.6815	0.0056
2	Immst	2	1	0.4535	0.0997	20.6724	< 0.0001
2	Immst	7	1	-2.4083	1.1271	4.5657	0.0326
2	Renter	2	1	-0.4294	0.0981	19.1394	< 0.0001
2	Urban	2	1	0.2503	0.1191	4.4183	0.0356
2	Jstat	2	1	0.6896	0.1255	30.1753	< 0.0001
2	Jstat	97	1	3.0628	0.4009	58.3666	< 0.0001
2	Age*Marst	2	1	-0.0059	0.0109	0.2900	0.5902
2	Age*Marst	3	1	0.0350	0.0151	5.4050	0.0201
3	Intercept		1	0.7870	0.3028	6.7549	0.0093
3	Age		1	0.0015	0.0050	0.0933	0.7600
3	Age^2		1	-0.0009	0.0003	9.4534	0.0021
3	Edlev	2:LM	1	0.5811	0.1948	8.8980	0.0029
3	Edlev	3:M	1	0.3530	0.1312	7.2331	0.0072
3	Edlev	4:H	1	0.3310	0.2954	1.2555	0.2625
3	Edlev	5:97	1	0.5468	0.5168	1.1197	0.2900
3	Marst	2	1	-0.2233	0.1392	2.5736	0.1087
3	Marst	3	1	-0.3433	0.1520	5.0980	0.0240
3	Immst	2	1	0.1940	0.1319	2.1637	0.1413
3	Immst	7	1	-1.2383	0.5140	5.8030	0.0160
3	Stud	2	1	-0.1817	0.2705	0.4510	0.5019
3	Stud	3	1	-0.0990	0.2328	0.1806	0.6708
3	Stud	7	1	-2.0878	0.5876	12.6248	0.0004
3	Renter	2	1	-0.2822	0.1080	6.8291	0.0090
3	Jstat	2	1	1.2685	0.1183	115.0402	< 0.0001
3	Jstat	97	1	2.8893	0.2505	133.0785	< 0.0001

Table C.4: Summary of model fits, LTF, years 4 and 5 (Ontario).

Year	Variable	Lev0	Lev1	DF	Estimate	StdErr	WaldChiSq	ProbChiSq
4	Intercept			1	1.0571	0.2845	13.8103	0.0002
4	Sex	2		1	-0.1876	0.1203	2.4320	0.1189
4	Age			1	-0.0127	0.0068	3.4534	0.0631
4	Age^2			1	-0.0010	0.0004	8.1895	0.0042
4	Edlev	2:LM		1	0.4497	0.1941	5.3697	0.0205
4	Edlev	3:M		1	0.1914	0.1327	2.0808	0.1492
4	Edlev	4:H		1	0.0911	0.2799	0.1060	0.7447
4	Edlev	5:97		1	-0.7373	0.1558	22.4028	< 0.0001
4	Marst	2		1	-0.5052	0.1783	8.0280	0.0046
4	Marst	3		1	-0.5817	0.2547	5.2177	0.0224
4	Stud	2		1	0.4585	0.2532	3.2797	0.0701
4	Stud	3		1	0.2569	0.2068	1.5425	0.2142
4	Stud	7		1	-1.7067	0.6906	6.1085	0.0135
4	Renter	2		1	-0.4608	0.1121	16.8994	< 0.0001
4	Jstat	2		1	0.8778	0.1320	44.2185	< 0.0001
4	Jstat	97		1	2.0960	0.2224	88.8241	< 0.0001
4	Sex*Marst	2	2	1	0.5226	0.1998	6.8451	0.0089
4	Sex*Marst	2	3	1	0.1746	0.3151	0.3072	0.5794
4	Age*Marst	2		1	0.0429	0.0129	10.9932	0.0009
4	Age*Marst	3		1	0.0475	0.0144	10.8937	0.0010
5	Intercept			1	1.0961	0.3624	9.1504	0.0025
5	Age			1	0.0027	0.0062	0.1913	0.6618
5	Age^2			1	-0.0012	0.0004	9.5384	0.0020
5	Edlev	2:LM		1	0.5088	0.2573	3.9113	0.0480
5	Edlev	3:M		1	0.2751	0.1781	2.3858	0.1224
5	Edlev	4:H		1	-0.0293	0.3499	0.0070	0.9333
5	Edlev	5:97		1	-0.4981	0.2143	5.4031	0.0201
5	Marst	2		1	-0.3437	0.1962	3.0684	0.0798
5	Marst	3		1	-0.4605	0.1952	5.5644	0.0183
5	Stud	2		1	0.4287	0.3260	1.7296	0.1885
5	Stud	3		1	0.2195	0.2775	0.6256	0.4290
5	Stud	7		1	-2.8124	0.7865	12.7874	0.0003
5	HHtype	2		1	-0.8435	0.3063	7.5858	0.0059
5	Urban	2		1	0.5423	0.1841	8.6754	0.0032
5	Jstat	2		1	1.1992	0.1623	54.5922	< 0.0001
5	Jstat	97		1	2.1647	0.3551	37.1680	< 0.0001

Table C.5: Summary of model fits, LTF, year 6 (Ontario).

				Events	Non Events		
Year	Variable	Level	DF	Estimate	StdErr	WaldChiSq	ProbChiSq
6	Intercept		1	1.4727	0.5098	8.3446	0.0039
6	Age		1	0.0004	0.0100	0.0016	0.9681
6	Age^2		1	-0.0015	0.0005	9.3771	0.0022
6	Marst	2	1	-0.2876	0.2658	1.1708	0.2792
6	Marst	3	1	-0.4928	0.2593	3.6129	0.0573
6	Stud	2	1	0.0154	0.4636	0.0011	0.9735
6	Stud	3	1	-0.3578	0.4095	0.7632	0.3823
6	Stud	7	1	-2.0590	0.6590	9.7620	0.0018
6	Famtype	2	1	0.3038	0.3049	0.9923	0.3192
6	Famtype	3	1	0.6674	0.2712	6.0575	0.0138
6	Famtype	4	1	0.2335	0.2581	0.8180	0.3658
6	Urban	2	1	0.3844	0.1862	4.2596	0.0390
6	Jstat	2	1	1.2500	0.1798	48.3089	< 0.0001
6	Jstat	97	1	2.5658	0.4836	28.1537	< 0.0001
6	Age*Marst	2	1	0.0016	0.0172	0.0084	0.9272
6	Age*Marst	3	1	0.0529	0.0185	8.1574	0.0043

Table C.6: Summary of model fits, LTF, years 2 and 3 (Quebec).

Year	Variable	Level	DF	Estimate	StdErr	WaldChiSq	ProbChiSq
2	Intercept		1	0.9405	0.2556	13.5341	0.0002
2	Age		1	0.0198	0.0045	19.5885	<0.0001
2	Age^2		1	-0.0021	0.0003	47.8731	<0.0001
2	Edlev	2:LM	1	0.1272	0.1826	0.4856	0.4859
2	Edlev	3:M	1	-0.0229	0.1172	0.0383	0.8448
2	Edlev	4:H	1	-0.0259	0.3147	0.0068	0.9344
2	Edlev	5:97	1	-1.1692	0.3168	13.6185	0.0002
2	Stud	2	1	0.3106	0.2606	1.4203	0.2334
2	Stud	3	1	0.3829	0.2120	3.2633	0.0708
2	Stud	7	1	-1.3455	0.6798	3.9172	0.0478
2	Jstat	2	1	0.4419	0.1318	11.2444	0.0008
2	Jstat	97	1	3.2764	0.3876	71.4563	<0.0001
3	Intercept		1	0.3879	0.4090	0.8995	0.3429
3	Age		1	0.0097	0.0053	3.3445	0.0674
3	Age^2		1	-0.0023	0.0003	45.1350	<0.0001
3	Immst	2	1	0.6013	0.2563	5.5060	0.0190
3	Immst	7	1	-0.9888	0.2884	11.7542	0.0006
3	Stud	2	1	0.1256	0.3238	0.1504	0.6981
3	Stud	3	1	0.3966	0.2726	2.1164	0.1457
3	Stud	7	1	-0.9102	0.5656	2.5896	0.1076
3	Famtype	2	1	-0.0322	0.2011	0.0256	0.8728
3	Famtype	3	1	0.3877	0.1811	4.5824	0.0323
3	Famtype	4	1	0.0663	0.2278	0.0847	0.7710
3	Hhtype	2	1	-0.6920	0.3191	4.7031	0.0301
3	Urban	2	1	-0.2923	0.1252	5.4457	0.0196
3	Jstat	2	1	0.8954	0.1416	40.0129	<0.0001
3	Jstat	97	1	3.3447	0.3433	94.8934	<0.0001

Table C.7: Summary of model fits, LTF, years 4 and 5 (Quebec).

Year	Variable	Lev0	Lev1	DF	Estimate	StdErr	WaldChiSq	ProbChiSq
4	Intercept			1	1.4171	0.2989	22.4831	<0.0001
4	Age			1	0.0130	0.0050	6.7896	0.0092
4	Age^2			1	-0.0017	0.0003	24.8819	<0.0001
4	Edlev	2:LM		1	0.3787	0.2460	2.3697	0.1237
4	Edlev	3:M		1	0.1594	0.1612	0.9771	0.3229
4	Edlev	4:H		1	0.0245	0.3937	0.0039	0.9505
4	Edlev	5:97		1	-0.9315	0.2236	17.3577	<0.0001
4	HHsz	2		1	-0.3314	0.2538	1.7048	0.1917
4	HHsz	3		1	0.0527	0.2712	0.0377	0.8460
4	HHsz	4		1	-0.2044	0.2608	0.6145	0.4331
4	HHsz	5		1	-0.1174	0.3132	0.1404	0.7079
4	Urban	2		1	-0.3381	0.4356	0.6022	0.4377
4	Jstat	2		1	0.7583	0.1563	23.5475	<0.0001
4	Jstat	97		1	2.6499	0.3482	57.9207	<0.0001
4	HHsz*Urban	2	2	1	1.0036	0.5257	3.6448	0.0562
4	HHsz*Urban	3	2	1	0.0125	0.5144	0.0006	0.9807
4	HHsz*Urban	4	2	1	1.1021	0.5429	4.1210	0.0424
4	HHsz*Urban	5	2	1	-0.1085	0.5389	0.0405	0.8405
5	Intercept			1	1.2667	0.4264	8.8266	0.0030
5	Age			1	0.0080	0.0059	1.8694	0.1715
5	Age^2			1	-0.0017	0.0004	19.0324	<0.0001
5	Immst	2		1	0.8767	0.2847	9.4839	0.0021
5	Immst	7		1	0.0854	0.3849	0.0492	0.8245
5	HHsz	2		1	-0.6666	0.3096	4.6354	0.0313
5	HHsz	3		1	-0.7280	0.3182	5.2322	0.0222
5	HHsz	4		1	-0.8162	0.3138	6.7665	0.0093
5	HHsz	5		1	-0.0584	0.3832	0.0232	0.8789
5	Jstat	2		1	1.0559	0.1908	30.6313	<0.0001
5	Jstat	97		1	2.2084	0.3886	32.2926	<0.0001

Table C.8: Summary of model fits, LTF, year 6 (Quebec).

				Events	Non Events		
Year	Variable	Level	DF	Estimate	StdErr	WaldChiSq	ProbChiSq
6	Intercept		1	1.8951	0.2929	41.8512	<0.0001
6	Age		1	0.0068	0.0064	1.1312	0.2875
6	Age^2		1	-0.0019	0.0004	18.8550	<0.0001
6	Edlev	2:LM	1	0.8016	0.3747	4.5761	0.0324
6	Edlev	3:M	1	0.4690	0.2145	4.7786	0.0288
6	Edlev	4:H	1	0.3496	0.5525	0.4004	0.5269
6	Edlev	5:97	1	-0.7048	0.3350	4.4265	0.0354
6	Renter	2	1	-0.4279	0.1904	5.0517	0.0246
6	Jstat	2	1	0.7438	0.2306	10.4023	0.0013
6	Jstat	97	1	3.5313	0.7463	22.3911	<0.0001

Table C.9: Deciles of risk of model fits, years 2 to 4 (Ontario).

			Events		Non Events	
Year	Group	Total	Observed	Expected	Observed	Expected
2	1	441	293	281.149	148	159.851
2	2	441	330	333.125	111	107.875
2	3	441	347	350.103	94	90.897
2	4	441	352	363.481	89	77.519
2	5	441	369	370.011	72	70.989
2	6	441	378	378.904	63	62.096
2	7	441	379	387.577	62	53.423
2	8	441	395	391.535	46	49.465
2	9	441	403	395.546	38	45.454
2	10	438	422	416.568	16	21.432
3	1	443	238	236.875	205	206.125
3	2	443	336	327.442	107	115.558
3	3	443	364	368.677	79	74.323
3	4	443	394	387.912	49	55.088
3	5	443	392	395.886	51	47.114
3	6	443	405	401.037	38	41.963
3	7	443	404	406.089	39	36.911
3	8	444	407	409.660	37	34.340
3	9	443	411	410.760	32	32.240
3	10	443	417	423.662	26	19.338
4	1	388	205	206.605	183	181.395
4	2	388	275	280.663	113	107.337
4	3	388	308	306.938	80	81.062
4	4	388	335	323.699	53	64.301
4	5	388	339	335.524	49	52.476
4	6	388	345	342.092	43	45.908
4	7	388	347	347.144	41	40.856
4	8	388	350	350.739	38	37.261
4	9	388	351	355.067	37	32.933
4	10	388	357	363.527	31	24.473

Table C.10: Deciles of risk of model fits, years 5 and 6 (Ontario).

			Events		Non Events	
Year	Group	Total	Observed	Expected	Observed	Expected
5	1	326	221	226.495	105	99.505
5	2	326	280	277.413	46	48.587
5	3	326	290	290.501	36	35.499
5	4	326	305	297.644	21	28.356
5	5	326	309	301.339	17	24.661
5	6	326	304	304.708	22	21.292
5	7	326	305	306.539	21	19.461
5	8	326	306	307.712	20	18.288
5	9	326	310	311.074	16	14.926
5	10	325	308	314.573	17	10.427
6	1	302	220	227.344	82	74.656
6	2	302	269	264.648	33	37.352
6	3	302	275	272.877	27	29.123
6	4	302	278	277.626	24	24.374
6	5	303	284	282.110	19	20.890
6	6	302	289	284.865	13	17.135
6	7	302	283	287.095	19	14.905
6	8	302	288	287.898	14	14.102
6	9	302	292	289.852	10	12.148
6	10	305	293	296.685	12	8.315

Table C.11: Deciles of risk of model fits, years 2 to 4 (Quebec)

			Events		Non Events	
Year	Group	Total	Observed	Expected	Observed	Expected
2	1	310	168	171.078	142	138.922
2	2	310	214	219.726	96	90.274
2	3	310	258	240.348	52	69.652
2	4	310	240	249.383	70	60.617
2	5	310	263	257.183	47	52.817
2	6	310	261	261.572	49	48.428
2	7	310	263	263.953	47	46.047
2	8	311	263	266.115	48	44.885
2	9	310	264	266.892	46	43.108
2	10	310	298	295.750	12	14.250
3	1	278	139	140.459	139	137.541
3	2	278	194	201.082	84	76.918
3	3	278	231	225.548	47	52.452
3	4	278	239	238.623	39	39.377
3	5	278	252	246.010	26	31.990
3	6	278	244	250.832	34	27.168
3	7	278	258	253.805	20	24.195
3	8	278	257	257.769	21	20.231
3	9	281	265	262.703	16	18.297
3	10	274	265	267.160	9	6.840
4	1	240	150	149.637	90	90.363
4	2	240	188	185.296	52	54.704
4	3	240	207	199.813	33	40.187
4	4	240	205	206.949	35	33.051
4	5	240	200	210.820	40	29.180
4	6	240	217	213.561	23	26.439
4	7	240	218	216.136	22	23.864
4	8	240	216	219.112	24	20.888
4	9	240	227	223.125	13	16.875
4	10	237	225	228.550	12	8.450

Table C.12: Deciles of risk of model fits, years 5 and 6 (Quebec).

			Events		Non Events	
Year	Group	Total	Observed	Expected	Observed	Expected
5	1	207	137	145.501	70	61.499
5	2	207	181	171.089	26	35.911
5	3	207	188	183.531	19	23.469
5	4	207	193	187.426	14	19.574
5	5	207	197	188.926	10	18.074
5	6	207	184	189.671	23	17.329
5	7	207	183	190.692	24	16.308
5	8	207	188	192.140	19	14.860
5	9	207	196	197.370	11	9.630
5	10	206	198	198.654	8	7.346
6	1	190	143	143.689	47	46.311
6	2	190	169	165.581	21	24.419
6	3	190	175	172.228	15	17.772
6	4	190	169	176.275	21	13.725
6	5	190	180	178.272	10	11.728
6	6	190	180	180.519	10	9.481
6	7	190	183	181.422	7	8.578
6	8	192	180	183.736	12	8.264
6	9	191	188	184.101	3	6.899
6	10	185	181	182.178	4	2.822

Table C.13: ChiSquare p-values of Hosmer and Lemeshow statistic by year (Ontario).

Year	ChiSq	DF	ProbChiSq
2	8.378	8	0.397
3	5.517	8	0.701
4	5.737	8	0.677
5	9.958	8	0.268
6	6.216	8	0.623

Table C.14: ChiSquare p-values of Hosmer and Lemeshow statistic by year (Quebec).

Year	ChiSq	DF	ProbChiSq
2	9.865	8	0.275
3	6.642	8	0.576
4	10.096	8	0.258
5	19.116	8	0.014
6	10.352	8	0.241

Appendix D

Summary of Estimation and Modelling from SLID

D.1 Kaplan-Meier estimates

Tables D.1 and D.2 give the results from applying the weighted Kaplan-Meier methods from chapter 5 on jobless spells from SLID individuals living in Ontario and Quebec in 1999, respectively. The estimates shown correspond to jobless spells that started in the years 1999 and 2000 and are associated to the times in which the estimated survival probability was closest to $0.1, 0.2, \ldots, 0.9$ with stardard errors in parenthesis. These tables show results when using unity weights, in the column labeled as "Unw". The columns "DES" and "COMB" show the results from using design weighted and combined weighted (design*IPC) methods.

D.2 Cox PH model analysis

Tables D.4 and D.5 show results from estimating parameters from the Cox PH model, which are approximated by a piecewise constant model (PC). First jobless spells from SLID that started in the year 2000 and belong to individuals that lived in Ontario in 1999 were used. The methods employed are UNWEIGHTED (unity weights), DESIGN,

Table D.1: Estimated survival probabilities, jobless spells in Ontario.

Year 1999			
	Unw	DES	COMB
Time	Est (SE)	Est (SE)	Est (SE,N.SE)
3	0.908 (0.015)	0.918 (0.016)	0.919 (0.014, 0.016)
7	0.805 (0.021)	0.827 (0.022)	0.831 (0.019, 0.022)
11	0.733 (0.024)	0.748 (0.027)	0.755 (0.024, 0.026)
19	0.596 (0.027)	0.595 (0.032)	0.609 (0.027, 0.032)
29	0.501 (0.028)	0.510 (0.034)	0.517 (0.029, 0.035)
35	0.399 (0.027)	0.398 (0.033)	0.405 (0.030, 0.035)
45	0.311 (0.026)	0.297 (0.032)	0.298 (0.026, 0.032)
70	0.202 (0.023)	0.186 (0.026)	0.188 (0.021, 0.027)
100	0.103 (0.018)	0.108 (0.022)	0.111 (0.018, 0.023)

Year 2000			
	Unw	DES	COMB
Time	Est (SE)	Est (SE)	Est (SE,N.SE)
3	0.911 (0.017)	0.894 (0.027)	0.898 (0.019, 0.026)
6	0.806 (0.024)	0.798 (0.032)	0.803 (0.023, 0.032)
10	0.693 (0.028)	0.700 (0.036)	0.712 (0.025, 0.035)
15	0.602 (0.030)	0.624 (0.037)	0.633 (0.028, 0.037)
21	0.486 (0.031)	0.502 (0.039)	0.520 (0.028, 0.040)
33	0.384 (0.030)	0.410 (0.041)	0.439 (0.029, 0.042)
39	0.292 (0.028)	0.317 (0.038)	0.339 (0.029, 0.041)
54	0.190 (0.025)	0.198 (0.032)	0.214 (0.027, 0.035)
108	0.086 (0.018)	0.083 (0.022)	0.098 (0.017, 0.026)

Table D.2: Estimated survival probabilities, jobless spells in Quebec.

Year 1999			
	Unw	DES	COMB
Time	Est (SE)	Est (SE)	Est (SE,N.SE)
3	0.906 (0.017)	0.898 (0.022)	0.898 (0.021, 0.022)
7	0.816 (0.022)	0.799 (0.030)	0.801 (0.027, 0.030)
13	0.699 (0.027)	0.657 (0.039)	0.665 (0.033, 0.038)
19	0.613 (0.029)	0.567 (0.042)	0.577 (0.036, 0.041)
26	0.503 (0.030)	0.484 (0.042)	0.488 (0.037, 0.041)
31	0.402 (0.029)	0.423 (0.041)	0.429 (0.037, 0.041)
39	0.304 (0.028)	0.327 (0.037)	0.334 (0.033, 0.037)
64	0.205 (0.025)	0.221 (0.033)	0.217 (0.027, 0.032)
119	0.112 (0.020)	0.146 (0.030)	0.133 (0.022, 0.028)

Year 2000			
	Unw	DES	COMB
Time	Est (SE)	Est (SE)	Est (SE,N.SE)
3	0.910 (0.020)	0.862 (0.040)	0.860 (0.038, 0.041)
4	0.800 (0.028)	0.769 (0.042)	0.768 (0.036, 0.043)
8	0.688 (0.032)	0.657 (0.050)	0.659 (0.042, 0.050)
19	0.563 (0.035)	0.537 (0.048)	0.547 (0.039, 0.048)
23	0.497 (0.035)	0.461 (0.054)	0.463 (0.047, 0.056)
30	0.388 (0.035)	0.383 (0.050)	0.380 (0.047, 0.054)
38	0.305 (0.033)	0.283 (0.044)	0.268 (0.040, 0.047)
53	0.218 (0.029)	0.199 (0.037)	0.181 (0.029, 0.037)
97	0.104 (0.022)	0.094 (0.025)	0.084 (0.019, 0.023)

which corresponds to the method based on Boudreau and Lawless [9] (results based on Binder [4] and Lin et al. [39] were very similar). The COMBINED method consists of variance estimates based on the proposed techniques in chapter 6.

Similarly, Tables D.6 and D.7, show the results obtained from sequences of jobless spells that started in 2000-2001, from SLID individuals that lived in Ontario in 1999.

Table D.3: Values of estimated constant hazards from first jobless spells starting in 2000 and sequences of jobless spells in 2000-2002.

First jobless spells starting in 2000.					
	$(b_{j-1}, b_j]$		Unweighted	Design	Combined
1	0.00	7.14	0.018	0.020	0.027
2	7.14	7.86	0.011	0.017	0.022
3	7.86	17.29	0.027	0.023	0.031
4	17.29	25.00	0.013	0.016	0.020
5	25.00	38.21	0.011	0.015	0.020
6	38.21	50.00	0.024	0.025	0.030
7	50.00	100.00	0.016	0.024	0.030
8	100.00	150.00	0.010	0.011	0.016
9	150.00	193.00	0.006	0.007	0.006

Sequences of jobless spells starting in 2000-2002.					
	$(b_{j-1}, b_j]$		Unweighted	Design	Combined
1	0.00	7.14	0.011	0.011	0.010
2	7.14	7.86	0.004	0.006	0.004
3	7.86	17.29	0.013	0.013	0.012
4	17.29	25.00	0.008	0.008	0.007
5	25.00	38.21	0.010	0.011	0.012
6	38.21	50.00	0.012	0.012	0.012
7	50.00	100.00	0.008	0.009	0.009
8	100.00	150.00	0.004	0.004	0.005
9	150.00	193.00	0.003	0.004	0.002

Table D.4: Unweighted estimation results, PC and Cox PH models. SLID first jobless spells starting in 2000.

	UNWEIGHTED			
	COX		PC	
	Estimate	SE*	Estimate	SE*
Age	-0.0125	0.0061	-0.0127	0.0063
Sex	0.1144	0.1691	0.1037	0.1733
Minority	0.2353	0.2651	0.2363	0.2699
Ei	-0.0826	0.2329	-0.1093	0.2374
Occup2	0.2605	0.3016	0.2666	0.3120
Occup3	-0.1902	0.3228	-0.2312	0.3327
Occup4	-0.2245	0.8201	-0.2973	0.8413
Occup5	0.1712	0.3175	0.1766	0.3233
Occup6	-0.1797	0.3068	-0.1970	0.3184
Occup99	-0.7660	0.3535	-0.7998	0.3671
Income.cat2	0.3511	0.2326	0.3735	0.2423
Income.cat3	0.4468	0.2652	0.4703	0.2740

* Based on robust variance estimates without strata, used for assessment of PC approximation.

Table D.5: Sampling design and combined with IPC weighted estimation results (PC and Cox PH models). SLID first jobless spells starting in 2000.

DESIGN

	COX		PC		
	Estimate	SE*	Estimate	SE*	SE.$B\&L$
Age	-0.0170	0.0068	-0.0177	0.0071	0.0071
Sex	0.0231	0.1787	0.0178	0.1849	0.1831
Minority	0.2241	0.2839	0.1909	0.2777	0.2779
Ei	-0.3295	0.2913	-0.3839	0.3022	0.3020
Occup2	0.1485	0.3309	0.1703	0.3443	0.3412
Occup3	-0.1759	0.3999	-0.2208	0.4149	0.4174
Occup4	-0.1391	0.8504	-0.2350	0.8992	0.9014
Occup5	0.0326	0.3580	0.0339	0.3706	0.3722
Occup6	-0.1834	0.3292	-0.1984	0.3478	0.3460
Occup99	-1.0991	0.3904	-1.1294	0.4121	0.4165
Income.cat2	0.4957	0.2435	0.5435	0.2557	0.2602
Income.cat3	0.7176	0.2724	0.7673	0.2819	0.2775

COMBINED

	COX		PC			
	Estimate	SE*	Estimate	SE*	SE.COMB	Naive
Age	-0.0225	0.0070	-0.0233	0.0074	0.0045	0.0074
Sex	-0.1481	0.2016	-0.1887	0.2177	0.1413	0.2149
Minority	0.1789	0.2675	0.1271	0.2602	0.1671	0.2576
Ei	-0.2497	0.2603	-0.2982	0.2701	0.2015	0.2699
Occup2	0.1168	0.3271	0.1274	0.3452	0.2853	0.3417
Occup3	-0.2475	0.3863	-0.3009	0.4055	0.3091	0.4074
Occup4	-0.5871	1.0138	-0.6450	1.0453	0.5573	1.0453
Occup5	-0.0501	0.3724	-0.0358	0.3878	0.2948	0.3898
Occup6	-0.2845	0.3426	-0.2853	0.3654	0.2856	0.3658
Occup99	-1.3375	0.4183	-1.3540	0.4479	0.3211	0.4522
Income.cat2	0.3051	0.2610	0.3687	0.2751	0.2044	0.2798
Income.cat3	0.6421	0.2628	0.7324	0.2764	0.1719	0.2707

* Based on robust variance estimates without strata, used for assessment of PC approximation.

Table D.6: Unweighted estimation results, PC and Cox PH models. SLID sequences of jobless spells starting in 2000-2002.

	UNWEIGHTED			
	COX		PC	
	Estimate	SE*	Estimate	SE*
Age	-0.0287	0.0045	-0.0289	0.0046
Sex	0.1664	0.0944	0.1610	0.0951
Minority	0.3041	0.1639	0.3080	0.1654
Ei	0.1028	0.1358	0.0981	0.1363
Occup2	0.4547	0.1799	0.4507	0.1792
Occup3	0.2648	0.1873	0.2529	0.1871
Occup4	-0.4203	0.3297	-0.4256	0.3310
Occup5	0.3186	0.1875	0.3216	0.1867
Occup6	0.1017	0.1894	0.0925	0.1886
Occup99	-0.1070	0.2298	-0.1099	0.2301
Income.cat2	0.5072	0.1201	0.5090	0.1210
Income.cat3	0.6559	0.1547	0.6572	0.1556
Quarts2	-0.1649	0.1315	-0.1670	0.1325
Quarts3	-0.0691	0.1209	-0.0749	0.1222
Quarts4	-0.0865	0.1294	-0.0907	0.1298
Order	0.2311	0.0984	0.2322	0.0995
Order:P.Dur	-0.0054	0.0040	-0.0053	0.0041

* Based on robust variance estimates without strata, used for assessment of PC approximation.

Table D.7: Sampling design weighted estimation results (PC and Cox PH models). SLID sequences of jobless spells starting in 2000-2002.

	DESIGN				
	COX		PC		
	Estimate	SE*	Estimate	SE*	SE.$B\&L$
Age	-0.0300	0.0056	-0.0305	0.0057	0.0057
Sex	0.0443	0.1173	0.0399	0.1174	0.1161
Minority	0.2914	0.1727	0.2908	0.1737	0.1740
Ei	-0.0888	0.1800	-0.0981	0.1819	0.1809
Occup2	0.3813	0.2495	0.3847	0.2491	0.2495
Occup3	0.4027	0.2487	0.3891	0.2487	0.2488
Occup4	-0.2758	0.3641	-0.2823	0.3670	0.3683
Occup5	0.2882	0.2555	0.2874	0.2544	0.2551
Occup6	0.1457	0.2542	0.1420	0.2538	0.2544
Occup99	-0.2144	0.3009	-0.2098	0.2992	0.3006
Income.cat2	0.5860	0.1470	0.5852	0.1462	0.1470
Income.cat3	0.8744	0.1718	0.8823	0.1723	0.1712
Quarts2	-0.2263	0.1562	-0.2269	0.1567	0.1570
Quarts3	-0.1168	0.1376	-0.1260	0.1382	0.1377
Quarts4	-0.1788	0.1495	-0.1928	0.1498	0.1497
Order	0.1625	0.1173	0.1600	0.1185	0.1185
Order:P.Dur	-0.0084	0.0055	-0.0085	0.0056	0.0055

* Based on robust variance estimates without strata, used for assessment of PC approximation.

Table D.8: Combined sampling design with IPC weighted estimation results (PC and Cox PH models). SLID sequences of jobless spells starting in 2000-2002.

	COMBINED					
	COX		PC			
	Estimate	SE*	Estimate	SE*	SE.COMB	Naive
Age	-0.0372	0.0060	-0.0381	0.0061	0.0049	0.0062
Sex	-0.0003	0.1213	-0.0044	0.1212	0.1100	0.1196
Minority	0.2853	0.1860	0.2862	0.1888	0.1592	0.1890
Ei	-0.0488	0.1944	-0.0535	0.2002	0.1727	0.1983
Occup2	0.5133	0.3212	0.5197	0.3276	0.2908	0.3280
Occup3	0.5969	0.3160	0.5882	0.3237	0.2800	0.3230
Occup4	-0.3236	0.4285	-0.2996	0.4431	0.4040	0.4448
Occup5	0.6050	0.3298	0.6221	0.3405	0.2867	0.3399
Occup6	0.2665	0.3261	0.2547	0.3338	0.2932	0.3344
Occup99	-0.0852	0.3708	-0.0902	0.3768	0.3368	0.3767
Income.cat2	0.5008	0.1551	0.4893	0.1541	0.1422	0.1542
Income.cat3	0.8566	0.1701	0.8514	0.1726	0.1513	0.1714
Quarts2	-0.2815	0.1584	-0.2875	0.1616	0.1404	0.1614
Quarts3	-0.2505	0.1473	-0.2515	0.1512	0.1311	0.1509
Quarts4	-0.2869	0.1613	-0.2770	0.1625	0.1455	0.1628
Order	0.1891	0.1156	0.1913	0.1158	0.1052	0.1156
Order:P.Dur	-0.0096	0.0050	-0.0095	0.0051	0.0047	0.0051

* Based on robust variance estimates without strata, used for assessment of PC approximation.

References

[1] P.K. Andersen, O. Borgan, R.D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer-Verlag, New York, 1993.

[2] P.K. Andersen, J.P Klein, K.M. Knudsen, and R. Tabanera y Palacios. Estimation of variance in cox's regression model with shared gamma frailties. *Biometrics*, 53(1):1475–1484, 1997.

[3] D.A. Binder. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51(1):279–92, 1983.

[4] D.A. Binder. Fitting cox's proportional hazards models from survey data. *Biometrika*, 79(1):139–47, 1992.

[5] D.A. Binder and G.R. Roberts. Can informative designs be ignorable? In *Survey Research Methods Section Newsletter, Issue 12*. American Statistical Association, 2001.

[6] H.P. Blossfeld and A. Hamerle. Using cox models to study multiepisode processes. *Sociological Methods and Research*, 17(4):432–448, 1989.

[7] H.P. Blossfeld, A. Hamerle, and Mayer K.U. *Event history analysis*. L. Erebaum Associates, Hillsdale, New Jersey, 1989.

[8] C. Boudreau. *Duration Data Analysis in Longitudinal Surveys*. PhD thesis, University of Waterloo, 2003.

[9] C. Boudreau and J.F. Lawless. Survival analysis based on the proportional hazards model and survey data. *The Canadian Journal of Statistics*, 34(2):203–216, 2006.

[10] M. Callegaro. Seam effects in longitudinal surveys. *Journal of Official Statistics*, 24:387–409, 2008.

[11] R.L. Chambers. Introduction to part a: Approaches to inference. In R.L. Chambers and C.J. Skinner, editors, *Analysis of Survey Data*. Wiley, 2003.

[12] R.L. Chambers and C.J. Skinner. *Analysis of survey data*. Wiley Series in Survey Methodology, UK, 2003.

[13] W.G. Cochran. *Sampling Techniques*. John Wiley and Sons, New York, 1977.

[14] D. Collet. *Modelling binary data*. Chapman and Hall/CRC, New York, second edition, 2003.

[15] R.J. Cook and J.F. Lawless. *The Statistical Analysis of Recurrent Events*. Wiley, New York, first edition, 2007.

[16] C. Cotton and P. Gilles. The seam effect in the slid. *SLID working paper No. 75F0002M. Statistics Canada, Ottawa, ON*, 1998.

[17] D.R. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B Methodological*, 34(1):187–220, 1972.

[18] D.R. Cox. Partial likelihood. *Biometrika*, 1(1):269–276, 1975.

[19] G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs. *Longitudinal Data Analysis*. Chapman and Hall/CRC Handbooks of Modern Statistical Methods., Chapman and Hall/CRC, Boca Raton., 2009.

[20] R. Folsom, L. LaVange, and R.L. Williams. *A probability sampling perspective on panel data analysis*. Kasprzyk et al., 1989.

[21] D. Galarneau and L.M. Stratychuk. After the layoff. *Perspectives, Statistics Canada*, 3(10):19–29, 2001.

[22] A. Gelman. Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2):153–164, 2007.

[23] A. Hamerle. Multiple spell regression models for duration data. *Applied statistics*, 38(1):127–138, 1989.

[24] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, New York, second edition, 2000.

[25] J.D. Kalbfleisch and R.L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, New York, second edition, 2002.

[26] G. Kalton, D.P. Miller, and J. Lepkowski. Analyzing spells of program participation in the sipp. Technical report, Survey Research Center, University of Michigan, 1992.

[27] M.C. Kenward and J. Carpenter. Multiple imputation: current perspective. *Statistical Methods in Medical Research*, 16:199–218, 2007.

[28] E.L. Korn and B.I. Graubard. *Analysis of Health Surveys*. John Wiley and Sons, New York, 1999.

[29] M.S. Kovacevic and G. Roberts. Modelling durations of multiple spells from longitudinal survey data. *Surve Methodology; Statistics Canada*, 33(1):13–22, 2007.

[30] S. LaRoche. Longitudinal and cross-sectional weighting of the survey of labour and income dynamics 2003. Technical report, Statistics Canada, Income Research Paper Series. Catalogue no. 75F0002MIE-007, 2007.

[31] J.F. Lawless. Event history analysis and longitudinal surveys. In R.L. Chambers and C.J. Skinner, editors, *Analysis of Survey Data*. Wiley, 2003.

[32] J.F. Lawless. *Statistical Models and Methods for Lifetime Data*. Wiley, New York, second edition, 2003.

[33] J.F. Lawless and D.Y.T. Fong. State duration models in clinical and observational studies. *Statistics in Medicine*, 18(1):2365–2376, 1999.

[34] J.F. Lawless, J.D. Kalbfleisch, and C.J. Wild. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society B*, 61:413–138, 1999.

[35] J.F. Lawless and Wigg M.B. Analysis of repeated failures or durations, with application to shunt failures for patients with paediatric hydrocephalus. *Applied Statistics*, 4(4):449–465, 2001.

[36] G. Lemaitre. Dealing with the seam effect problem for the survey of labor and income dynamics. *SLID Research Paper 92-05. Statistics Canada, Ottawa, ON*, 1992.

[37] D.Y. Lin. Cox regression analysis of multivariate failure time data: a marginal approach. *Statistics in Medicine*, 14(1):2233–2247, 1994.

[38] D.Y. Lin. On fitting cox's proportional hazards models to survey data. *Biometrika*, 87:37–47, 2000.

[39] D.Y. Lin, W. Sun, and Z. Ying. Nonparametric estimation of the gap time distribution for serial events with censored data. *Biometrics*, 86(1):59–70, 1999.

[40] D.Y. Lin and L.J. Wei. The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association*, 84:1074–1078, 1989.

[41] R.J.A Little. Survey inference with weights for differential sample selection or non-response. In *Proceedings of the Section on Survey Research Methods*, pages 62 – 69. American Statistical Association, 1989.

[42] R.J.A. Little. To model or not to model? competing modes of inference for finite population sampling. *American Statistical Association*, 99(466):546 –556, 2004.

[43] Little L.J.A. and Rubin D.B. *Statistical Analysis of Missing Data*. Wiley, New York, 2002.

[44] M. E. Miller, Ten Have T.R., B.A. Reboussin, K.K. Lohman, and W. J. Rejeski. A marginal model for analyzing discrete outcomes from longitudinal surveys with outcomes subject to multiple cause nonnesponse. *Journal of the American Statistical Association*, 96(455):844–857, 2001.

[45] C.G. Moertel, T.R. Fleming, and J.S. McDonald. Levamisole and fluorouracil for adjuvant therapy of restricted colon carcinoma. *The New England Journal of Medicine*, 322(1):352–358, 1990.

[46] D. Pfeffermann. The role of sampling weights when modeling survey data. *International Statistical Review*, 61(2):317–337, 1993.

[47] I. Plewis. Non-response in a birth cohort study: The case of the millennium cohort study. *International Journal of Social Research Methodology*, 10(5):325–334, 2007.

[48] J.S. Preisser, K.K. Lohman, and P.J. Rathouz. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, 21(1):3035–3054, 2002.

[49] J.M. Robins. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the Biopharmaceutical Section*. American Statistical Association, 1993.

[50] J.M. Robins and Finkelstein D.M. Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics*, 56(1):779–788, 2000.

[51] J.M. Robins, A. Rotnitzky, and L.P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(1):106–121, 1995.

[52] A. Rotnitzky, J.M. Robins, and D.O. Scharfstein. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American statistical association*, 93(1):1321–1339, 1998.

[53] D.B. Rubin. Inference and missing data. *Biometrika*, 63:581–92, 1976.

[54] D.B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91:473–489, 1996.

[55] G.A. Satten, S. Datta, and J. Robins. Estimating the marginal survival function in the presence of time dependent covariates. *Statistics and Probability Letters*, 54(4):397–403, 2001.

[56] D.O. Scharfstein and J.M. Robins. Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, 89:617–634, 2002.

[57] T.M. Therneau and P.M. Grambsch. *Modeling survival data: extending the Cox model*. Springer-Verlag, New York, first edition, 2000.

[58] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

[59] G.Y. Yi and R.J. Cook. Marginal methods for incomplete longitudinal data arising in clusters. *Journal of the American statistical association*, 97(460):1071–1080, 2005.

[60] G.Y. Yi and M.E. Thompson. Marginal and association regression models for longitudinal binary data with drop-outs: a likelihood-based approach. *The Canadian Journal of Statistics*, 33(1):3–20, 2005.