View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by University of Waterloo's Institutional Repository

Estimation of Finite Population Duration Distributions from Longitudinal Survey Panels with Intermittent Followup

DAGMAR M. HAJDUCEK

Department of Psychology,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

E-mail: hajducek@gmail.com

J. F. LAWLESS

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

E-mail: jlawless@uwaterloo.ca

Summary

We consider survival or duration times associated with spells (sojourns in some state) or events experienced by individuals in a population over a specified time period. Duration distributions can be estimated from data recorded during followup of panel members in longitudinal surveys, but adjustments for the sample design, population structure and losses to followup are typically required. We provided weighted Kaplan-Meier estimates that allow for these features and, in particular, adjust for dependent loss to followup through the use of inverse probability of censoring weights.

Keywords: Dropout, Inverse probability weights, Loss to followup, Spell durations, Survival distributions, Weighted Kaplan-Meier estimates

The final publication: "Hajducek, D.M. and Lawless, J.F. (2013). Estimation of finite population duration distributions from longitudinal survey panels with intermittent followup. *Lifetime Data Analysis*, 19 (3), 371–392" is available at Springer via DOI: 10.1007/s10985-012-9241-5.

1 INTRODUCTION

Models in which individuals may spend time in various states are used in economics, medicine, sociology and other areas. For example, in this paper we consider employment histories in which a person may be unemployed, employed, or not in the labor force at any given time. We will use the term spell to denote a period in which an individual is in some specific state; this terminology is common in economics and the social sciences (e.g. see Kovacevic and Roberts 2007; Pyy-Martikainen and Rendtel 2009).

Survival or duration distributions associated with spells experienced by individuals in a finite population are of considerable interest as descriptive quantities. For example, Figure 1 shows weighted Kaplan-Meier (KM) survival function estimates, obtained using methodology developed in this paper,

for the durations of jobless spells starting in Ontario and Quebec, respectively, in 1999 and 2000, for persons residing in Ontario and Quebec in 1999. These are based on data from Statistics Canada's Survey of Labour and Income Dynamics (SLID) and they estimate finite population duration distributions, defined as follows. Let N spells with durations $Y_1, Y_2, \ldots, Y_N$ be experienced by individuals in a specified population P, over a specified period of time. Then

$$S_P(y) = \frac{1}{N} \sum_{j=1}^{N} I\left(Y_j \geq y\right) \tag{1}$$

gives the population duration distribution. Note that a given individual may have more than one spell (or no spells), but the interest here is on the population aggregate number of spells and their durations and not on the dynamics of jobless spells at the person level. One may also be interested in distributions for the time to a specific event. For example, the U.K. Millennium Cohort Study (MCS) follows a cohort sampled from children born in 2000 and 2001 (Plewis 2007), and variables such as the age at which a child reaches a developmental milestone are of interest. In this case each individual has just one Y_j.

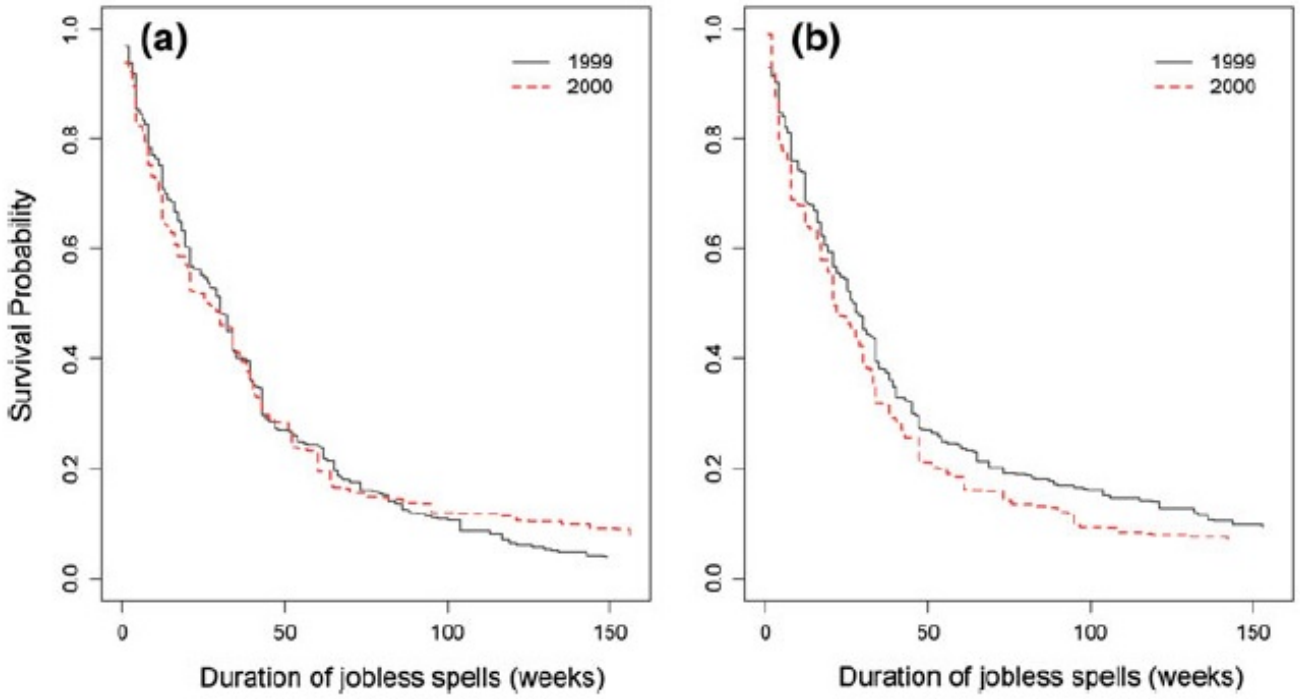

Figure 1: Weighted KM estimates for jobless spells starting in 1999 and 2000, from 1999 residents in (a) Ontario, 359 and 270 spells; (b) Quebec, 311 and 211 spells

The purpose of this paper is to consider estimation of finite population distributions (1) based on data from cohorts or longitudinal surveys in which a small proportion of the population have been selected for inclusion in a panel. Such data have features that standard survival analysis methods (e.g. Lawless 2003a) do not address. In particular, the populations considered are typically heterogeneous, and study individuals are selected according to a survey design involving stratification and clustering. In addition, data on duration variables are collected at intermittent visits or interviews scheduled over a long period of time, and individuals may become lost to followup (LTF) before the final visit. For example, in SLID the panel members are seen annually for six years, and LTF by year 6 is typically in the 20-30% range. In

the MCS, children were seen over the first seven years at ages 9 months, 3, 5 and 7 years. Approximately 42% of the initial cohort were LTF by age 3 (Plewis 2007). Loss to followup is a serious issue when estimating distributions (1), because it may depend on covariates or other factors that are also related to spell durations Y.

A number of authors have considered estimation of (1) from survey data by developing KM estimates incorporating survey sampling weights (Folsom et al. 1981; Kalton et al. 1992; Williams 1995). Korn and Graubard (1999) consider alternative estimates based on the combination of stratum-specific survivor functions. Aside from Lawless (2003b), previous work has disregarded the possibility of dependent loss to followup, which arises because LTF is related to covariates or previous responses that are also related to spell duration. This is a serious deficiency, since LTF is almost always related to spell duration when adjustment for covariates is not made. We remedy this by developing weighted KM estimators that combine survey design weights with inverse probability of censoring weights (IPCW), as proposed by Robins (1993) and Satten et al. (2001). We provide variance estimates that account for the use of a complex sampling design involving stratification and clustering as well as the estimation of loss to followup probabilities. In addition, we discuss alternative estimators of (1) based on the use of regression or "augmented" estimation in a finite population context.

The observational framework we consider is as follows. Individuals selected for a panel are seen at times $t = 0, 1, 2, \ldots, M$ over a period $[0, M]$. At the initial visit $(t = 0)$, baseline information about an individual is collected; this may or may not include details of events or spells that started before $t = 0$. At visit t $(t = 1, \ldots, M)$, data $D_i(t)$ on the time period $(t - 1, t]$ are collected. It is assumed that in the case of spells, their exact start and end times can be obtained; cases with measurement error are discussed briefly in the final section. An important consequence of the intermittent data collection and LTF is that different weights should be applied to the data for different time periods. Since a spell or a period involving a time to an event typically overlaps more than one interval $(t - 1, t]$, this leads to a weighted KM estimator with time-varying weights.

The remainder of the paper is organized as follows. Section 2 discusses duration distributions in a finite population context, and Section 3 develops weighted KM estimation in the longitudinal survey setting. Section 4 gives results of simulation studies, comparing our estimates and previous design-weighted estimates. Section 5 applies the methodology to the estimation of jobless spell duration distributions, based on data from SLID. Section 6 makes some concluding remarks, including discussion of competing risks, alternative approaches to estimation based on regression and augmentation (e.g. Van der Laan et al., 2002) and missing or mis-measured data.

2 Duration Distributions in Finite Populations

Duration variables are associated with the durations of spells, or with times to or between events. For the sake of exposition we will treat a duration as the length of a spell or episode spent in some specific state. Each spell has a calendar start date U and end date V and $Y = V - U$ is the duration. In some contexts each individual in a finite population may have a latent duration time associated with them. Such is the case for the time to a developmental milestone for children in scope for the MCS, for example. Often, however, only some individuals in a population or sample have a spell and an associated duration; for example, in SLID only some individuals experience a jobless spell over a given period of time. Moreover, in some contexts such as jobless spells, an individual may have more than one spell and corresponding durations in a given time period. For example, the estimates in Figure 1 involve some individuals who had more than one jobless spell in the years indicated.

Thus, it is necessary to be precise in defining a duration distribution (1). We assume that in the

population of interest, the durations of spells that start during a specified time period are the focus of estimation, and we will use the following notation. Let $m_i \geq 0$ denote the number of spells beginning in the time period for individual i in the population P and if $m_i > 0$ let Y_{ij} $(j = 1, \ldots, m_i)$ be the duration of the jth spell. We then re-express (1) as

$$S_P(y) = \frac{1}{N} \sum_{i \in P} \sum_{j=1}^{m_i} I\left(Y_{ij} \geq y\right), \tag{2}$$

where $N = \sum_{i \in P} m_i$ is the total number of spells across the population and it is understood that terms in (2) equal 0 for persons with $m_i = 0$. Although the spells in question always refer to some specified calendar time period, we will not indicate this explicitly except where necessary. In addition, we will in all cases assume an upper limit has been placed on y in (2). For example, if (2) represents spells that begin in the first two years of a six-year longitudinal survey, then we impose a limit of 4 years or less on y, in order to make (2) estimable. Finally, note that no assumptions are made, or needed, concerning relationships between multiple spells for individuals or for variations in spell lengths over time.We reiterate that the object of interest here is a finite population quantity, given by (1) and (2), so it is a descriptive measure. If one were interested in explanatory analysis involving covariates, then consideration of within and between individual variation in spells would be important.

Although (2) is a finite population quantity, for prospective longitudinal surveys its components are latent at the time a sample is selected, and the processes generating the spells and durations, as well as losses to followup, are random. It is sometimes useful to consider a conceptual super-population survivor function $S(y)$ associated with (2), as follows. Let the size N^* of P become arbitrarily large, while keeping the same basic structure, and assume that as $N^* \to \infty$, $S_P(y)$ converges in probability to a limit distribution $S(y)$. That is,

$$S(y) = \text{plim } S_P(y) = \text{plim} \left(\frac{N^*}{N}\right) \text{plim} \left(\frac{1}{N^*} \sum_{i \epsilon P} \sum_{j=1}^{m_i} I\left(Y_{ij} \geq y\right)\right). \tag{3}$$

Here, N and the durations are random variables in the process generating the population and durations.

Finite populations are usually stratified for sampling, and there may be significant differences in the duration distributions across strata. Differences across large strata (e.g. differences in jobless spell duration distributions across certain provinces or states) are normally of substantive interest, so that separate estimates are considered for each stratum. We assume for the development here that some portion of the population has been specified and we consider estimation for it. Survey weights are used to adjust for variation in sampling rates across design strata, and stratification will be recognized in obtaining variance estimates.

3 Weighted Kaplan-Meier Estimation

3.1 Weighted Estimates

We consider a discrete time scale (e.g. days, weeks) for the timing of events, spells and durations. We refer to $S(y)$ of (3) as a probability distribution and for simplicity develop estimation in terms of it, but the estimates also apply to the finite population distribution (2). We define for $S(y)$ the associated hazard function

$$h(y) = \Pr\left(Y = y | Y \geq y\right) = \frac{f(y)}{S(y)}, \quad y = 1, 2, 3, \ldots, T \tag{4}$$

where $f(y) = S(y) - S(y+1)$ and similarly define $h_P(y)$. In practice we place some upper limit on the range of y, here taken as $T+1$; by definition we then set $h(T+1) = 1$.

We assume as in Section 2 that an individual may have $m_i \geq 0$ spells, and if $m_i > 0$ for individual i we let U_{ij} and V_{ij} denote the start and end date for their jth spell, with $Y_{ij} = V_{ij} - U_{ij}$ $(j = 1, \ldots, m_i)$. Multiple spells for the same individual are not assumed independent in the superpopulation model. The observational framework described in Section 1 involves the collection of data $D_i(t)$ on spell durations for panel (sample) members, for the calendar time interval $(t-1, t]$ for $t = 1, \ldots, M$, along with baseline data $D_i(0)$ collected at $t = 0$. To allow for premature LTF, we denote $C_i \in \{1, \ldots, M\}$ as the last visit at which panel individual i is seen; we consider only persons seen at $t = 0$, though sometimes there is non-response at this initial visit (e.g. Plewis 2007, Pyy-Martikainen and Rendtel 2008). We also define the indicator variables

$$R_{it} = I(\text{individual } i \text{ is seen at time } t) = I(C_i \geq t) \tag{5}$$

for the individuals, denoted for convenience as $i = 1, 2, \ldots, n$, selected for the panel in question.

The two key issues that we address are that (i) panel individuals are selected according to a survey design, in which the probability individual i is selected is π_i, and (ii) the probability a panel individual becomes LTF at a specific visit may depend on their previous event and duration history, as well as covariates. These features can lead to substantial bias in naive (unweighted) KM estimators. We now consider design and IPC weights, intended to produce consistent estimates. We apply the IPCW framework of Robins et al. (1995) to estimation of $S(y)$ from the duration data $D_i(t)$, $0 \leq t \leq C_i$, for each panel member. Such weighted KM estimation has been considered previously by Robins (1993), Robins and Finkelstein (2000), Satten et al. (2001) and others, but the details of the present development are somewhat different because of the complex sampling and intermittent data collection. Lawless (2003b) noted the desirability of using both design and IPC weights but did not develop variance estimates or explore the properties of the estimates.

We remark that an alternative way to estimate (2) or (3) is via regression models for Y. Such models are also of interest in explanatory analytical studies of spell durations (e.g. Hajducek and Lawless 2012; Kovacevic and Roberts 2007). The approach taken here is simpler to implement when estimation of (2) or (3) is the objective, and is consistent with related work on design-weighted and on IPCW Kaplan-Meier estimation. We discuss regression and related estimation methods briefly in Section 6.

The IPCW framework of Robins et al. (1995) applies when covariate vectors $Z_i^c(t)$ can be identified for $t = 1, \ldots, M$ such that R_{it} is independent of the data $D_i(s)$, $s \geq t$, given $Z_i^c(t)$. We make this important assumption, where $Z_i^c(t)$ can depend only on covariates or previous spell history up to time $t-1$; this assures that data at time t and later which are missing due to an individual being LTF at t are missing at random (MAR) in the terminology of Rubin (1976). We further denote

$$p_{it} = \Pr\left(R_{it} = 1 | Z_i^c(t)\right) = \Pr\left(C_i \geq t | Z_i^c(t)\right). \tag{6}$$

The p_{it} are unknown, and the IPCW approach assumes satisfactorily specified parametric models, $p_{it}(\alpha_t)$. As do many authors (e.g. Miller et al. 2001), we use logistic regression models,

$$\begin{aligned} \text{logit}\left(\lambda_{it}(\alpha_t)\right) &= \text{logit}\left\{\Pr(R_{it} = 1 | R_{i,t-1} = 1, Z_i^c(t))\right\} \\ &= \alpha_t' Z_i^c(t), \end{aligned} \tag{7}$$

where $\text{logit}(x) = \log(x/(1-x))$ and α_t is a vector of regression coefficients. By convention we write all vectors in column form. Note that $p_{it}(\alpha) = \lambda_{i1}(\alpha) \ldots \lambda_{it}(\alpha)$, where for notational convenience we let

α denote $(\alpha_1', \ldots, \alpha_M')'$. The model in (7) can be fitted with standard logistic regression or generalized linear model software to give maximum likelihood estimates $\hat{\alpha}_t$ and estimated probabilities $\hat{p}_{it} = p_{it}(\hat{\alpha}_t)$. Sufficient variables should be included in $Z_i^c(t)$ to make $D_i(t)$ conditionally independent of R_{it}, although it is impossible to verify this on the basis of the data obtained. Since we consider estimation of marginal duration distributions $S(y)$ without conditioning on any covariates, $Z_i^c(t)$ should include, for example, terms to reflect strata across which both LTF and duration distributions vary.

To obtain weighted KM estimators, we define interval-dependent indicators

$$\begin{aligned} d_{ijt}(y) &= I\left(Y_{ij} = y, t-1 < u_{ij} + y \leq t\right) R_{it}, \\ \delta_{ijt}(y) &= I\left(Y_{ij} \geq y, t-1 < u_{ij} + y \leq t\right) R_{it} \end{aligned}$$

for $t = 1, \ldots, M$ and $y = 1, \ldots, T$. For individual i we now consider the following estimating function for $h(y)$:

$$U_i^W(y) = \sum_{j=1}^{m_i} w_{ij}(y) \left[d_{ij}(y) - h(y)\right], \tag{8}$$

where

$$w_{ij}(y) = \sum_{t=1}^{M} \frac{\delta_{ijt}(y)}{\pi_i p_{it}(\alpha)}, \qquad d_{ij}(y) = \sum_{t=1}^{M} d_{ijt}(y). \tag{9}$$

It is shown in the Appendix that under the LTF assumptions above, $E\left\{R_i U_i^W(y)\right\} = 0$ for a random member of the population, where $R_i = I\big(\text{person } i \text{ is in the sample}\big)$. Thus the solution of

$$U^W(y) = \sum_{i=1}^{n} U_i^W(y) = 0, \qquad y = 1, \ldots, T \tag{10}$$

gives a consistent estimator of $h(y)$ and of $h_P(y)$. This is

$$\hat{h}(y) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \hat{w}_{ij}(y) d_{ij}(y) \Big/ \sum_{i=1}^{n} \sum_{j=1}^{m_i} \hat{w}_{ij}(y), \tag{11}$$

where $\hat{w}_{ij}(y)$ is given by (9) with α replaced by a consistent estimator $\hat{\alpha}$. The weighted KM estimator of $S(y)$ is then given by

$$\hat{S}(y) = \prod_{s=1}^{y-1} \left(1 - \hat{h}(s)\right), \qquad y = 1, 2, \ldots, T. \tag{12}$$

The IPCW estimation theory relies on a correctly specified model for loss to followup. Of course, no model is ever exactly "true" and in practice we use models that approximate a true process satisfactorily. Mild misspecification of the censoring process has little adverse effect but in practice we should take care to consider all variables that might be related to both durations and LTF, and to check the adequacy of fitted models. This is discussed in Section 5 and in references such as Miller et al. (2001), where it is noted that well established methods for checking the models used here exist. We also recommend examining the weights associated with each observation, across the different intervals $(t-1, t]$, and either trimming any very large (relative to the average) weights or assessing the effect of dropping that individual on estimates. Finally, we note that nonparametric approaches to deal with dependent censoring have been proposed (e.g. Stitelman and van der Laan 2010). We comment on other approaches in Section 6, but remark here that it is unclear how to adapt these methods to our exact setting.

3.2 VARIANCE ESTIMATION

Variance estimation for $\hat{\theta} = \left[\hat{h}(1), \ldots, \hat{h}(T)\right]'$ can be based on asymptotic theory for estimating functions (e.g. White 1982). We want to recognize the sampling design, which involves stratification and possible within-cluster association, and so we extend the notation of the preceding section along the lines of Miller et al. (2001), Kovacevic and Roberts (2007) and others. In particular, let us assume that individuals are sampled within R strata, with K_r clusters (primary sampling units) of sizes $n^*_{rk} (k = 1, \ldots, K_r)$ selected within stratum r $(r = 1, \ldots, R)$. We then let (r, k, i) indicate individual i within cluster k from stratum r, and rewrite $p_{it}(\alpha)$ in (9) as $p_{rkit}(\alpha)$, $\delta_{ijt}(y)$ as $\delta_{rkijt}(y)$, and so on.

The solution of the estimating equation (10), based on (8) with $\hat{w}_{ij}(y)$ replacing $w_{ij}(y)$, can be viewed as arising from the simultaneous solution of estimating equations for θ and α, where $\theta = [h(1), \ldots, h(T)]'$:

$$U(\theta, \alpha) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i=1}^{n^*_{rk}} U_{rki}(\theta, \alpha) = 0 \tag{13}$$

$$G(\alpha) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i=1}^{n^*_{rk}} G_{rki}(\alpha) = 0. \tag{14}$$

The terms in (13) are given by (8) as

$$U_{rki}(\theta_y, \alpha) = I\left(m_{rki} > 0\right) \sum_{j=1}^{m_{rki}} w_{rkij}(y) \left[d_{rkij}(y) - h(y)\right], \quad y = 1, \ldots, T \tag{15}$$

with $U(\theta, \alpha)$ a $T \times 1$ vector $(U(\theta_1, \alpha), \ldots, U(\theta_T, \alpha))'$, where $\theta_y = h(y)$ $(y = 1, \ldots, T)$. The estimating function $G(\alpha)$ in (14) is a $q \times 1$ vector, where $q = q_1 + \ldots + q_M$ and q_t is the dimension of α_t in (7). Its components come from the separate logistic regression log likelihood functions for $\alpha_1, \ldots, \alpha_M$. For α_t this is

$$\ell_t(\alpha_t) = \sum_{r=1}^{R} \sum_{k=1}^{k_r} \sum_{i=1}^{n^*_{rk}} R_{rki,t-1} \log\left\{\lambda_{rkit}(\alpha_t)^{R_{rkit}} \left[1 - \lambda_{rkit}(\alpha_t)\right]^{1-R_{rkit}}\right\}.$$

With $\lambda_{rkit}(\alpha_t)$ given by the logistic regression specification (7), the likelihood estimating function for α_t is the $q_t \times 1$ vector $G^{(t)}(\alpha_t) = \partial \ell_t(\alpha_t)/\partial \alpha_t$, which is:

$$\begin{aligned} G^{(t)}(\alpha_t) &= \sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i=1}^{n^*_{rk}} R_{rki,t-1} Z^c_{rki}(t) \left\{R_{rkit} - \lambda_{rkit}(\alpha_t)\right\} \\ &= \sum_{r=1}^{R} \sum_{k=1}^{K_r} \sum_{i=1}^{n^*_{rk}} G^{(t)}_{rki}(\alpha_t). \end{aligned} \tag{16}$$

The terms in (14) are then

$$G_{rki}(\alpha) = \left(G^{(1)}_{rki}(\alpha_1)', \ldots, G^{(M)}_{rki}(\alpha_M)'\right)'. \tag{17}$$

From an application of results in Robins et al. (1995) or White (1982), a consistent estimator of the asymptotic covariance matrix for $\hat{\theta}$ is given by

$$\widehat{Var}(\hat{\theta}) = A_{11}\left(\hat{\theta}, \hat{\alpha}\right)^{-1} \left\{B_{11}\left(\hat{\theta}, \hat{\alpha}\right) - B_{12}\left(\hat{\theta}, \hat{\alpha}\right) B_{22}(\hat{\alpha})^{-1} B_{21}\left(\hat{\theta}, \hat{\alpha}\right)\right\} A_{11}\left(\hat{\theta}, \hat{\alpha}\right)^{-1}, \tag{18}$$

where $A_{11}(\theta,\alpha)$, $B_{11}(\theta,\alpha)$, $B_{12}(\theta,\alpha)$, $B_{21}(\theta,\alpha)$ and $B_{22}(\alpha)$ are as follows; details are given in the Appendix. For $r = 1, 2, \ldots, R$ define

$$\begin{aligned}\bar{U}_r(\theta,\alpha) &= \frac{1}{K_r}\sum_{k=1}^{K_r}\sum_{i=1}^{n^*_{rk}} U_{rki}(\theta,\alpha) = \frac{1}{K_r}\sum_{k=1}^{K_r} U_{rk.}(\theta,\alpha)\\ \bar{G}_r(\alpha) &= \frac{1}{K_r}\sum_{k=1}^{K_r}\sum_{i=1}^{n^*_{rk}} G_{rki}(\alpha) = \frac{1}{K_r}\sum_{k=1}^{K_r} G_{rk,}(\alpha),\end{aligned}$$

where dot subscripts are used to denote summation.

Then, we have

$$B_{11}(\theta,\alpha) = \sum_{r=1}^{R}\frac{K_r}{K_r-1}\sum_{k=1}^{K_r}\left[U_{rk.}(\theta,\alpha)-\bar{U}_r(\theta,\alpha)\right]\left[U_{rk.}(\theta,\alpha)-\bar{U}_r(\theta,\alpha)\right]' \tag{19}$$

$$B_{12}(\theta,\alpha) = B_{21}(\theta,\alpha)' = \sum_{r=1}^{R}\frac{K_r}{K_r-1}\sum_{k=1}^{K_r}\left[U_{rk.}(\theta,\alpha)-\bar{U}_r(\theta,\alpha)\right]\left[G_{rk.}(\alpha)-\bar{G}_r(\alpha)\right]' \tag{20}$$

$$B_{22}(\alpha) = \sum_{r=1}^{R}\frac{K_r}{K_r-1}\sum_{k=1}^{K_r}\left[G_{rk.}(\alpha)-\bar{G}_r(\alpha)\right]\left[G_{rk.}(\alpha)-\bar{G}_r(\alpha)\right]' \tag{21}$$

and $A_{11}(\theta,\alpha)$ is a diagonal $T \times T$ matrix with diagonal entries

$$A_{11}(\theta,\alpha)_{yy} = \sum_{r=1}^{R}\sum_{k=1}^{K_r}\sum_{i=1}^{n^*_{rk}} I\left(m_{rki}>0\right)\sum_{j=1}^{m_{rki}} w_{rkij}(y), \qquad y = 1,\ldots,T \tag{22}$$

The dimensions of $B_{11}(\theta,\alpha)$, $B_{12}(\theta,\alpha)$ and $B_{22}(\alpha)$ are $T \times T$, $T \times q$ and $q \times q$, respectively. The matrices (19) - (21) estimate $Var\{U(\theta,\alpha)\}$, $Cov\{U(\theta,\alpha), G(\alpha)\}$ and $Var\{G(\alpha)\}$ and have been defined to reflect possible stratum effects and the use of a stratified sampling plan. We reiterate, however, that if LTF probabilities vary across strata then stratum effects should be incorporated in the covariates $Z^c(t)$ for the LTF models (7). Alternative variance estimates to (18) are discussed in the Appendix.

An asymptotic variance estimate for $\hat{S}(y)$ in (12) is given by a straightforward application of the delta theorem (Lawless 2003a, Appendix B.1), leading to

$$\widehat{Var}\left\{\hat{S}(y)\right\} = \hat{S}(y)^2\sum_{s=1}^{y-1}\sum_{s'=1}^{y-1}\frac{\widehat{Cov}\left[\hat{h}(s),\hat{h}(s')\right]}{\left[1-\hat{h}(s)\right]\left[1-\hat{h}(s')\right]}, \tag{23}$$

where $\widehat{Cov}[\hat{h}(s),\hat{h}(s')]$ is the (s,s') element of (18).

4 Simulation Studies

We present here simulation results that demonstrate the properties of combined design - IPC weights for survivor function estimation, and the inadequacy of using only design weights when LTF is duration-related. We constructed several finite populations of size $N = 100,000$, in which the individuals experience alternating "not jobless" and "jobless" spells over a 6-year period. Our objective is to estimate the

finite population survivor function (1) for the durations of jobless spells. The population is constructed with 10 strata of equal sizes (10,000), with jobless spell durations varying across strata as described below. Estimation of $S_P(y)$ in (1) is based on panels obtained by simple random samples of sizes 100, 110, 120, ..., 190 drawn from strata $1, 2, 3, \ldots, 10$ respectively, giving a total sample of $n = 1,450$ individuals. Individuals in the panel are seen annually and are subject to a LTF process, specified below. The performance of estimators discussed below is with respect to 1000 independent panels drawn by the stratified random sampling plan described above.

Table 1: Scenarios for jobless spell durations, measured in weeks

	EV	ρ	β_1	β_2
I	0.5	0.3	0.226	0.226
II	0.5	0.0	0.365	0.000
III	0.3	0.3	0.247	0.000
IV	0.3	0.3	0.1535	0.1535

For simplicity we consider two covariates (X_{1i}, X_{2i}) per individual, with values in the population generated from a bivariate normal distribution with both means 0, variances 1, and correlation ρ. For convenience we start each individual with a "not jobless" spell. Not jobless spells and jobless spells have durations that are mutually independent, given X_1 and X_2, and they are generated for the finite population of individuals as follows.

(i) Durations Y_{rij} for the jth jobless spell for person i in stratum r ($r = 1, \ldots, 10$; $i = 1, \ldots, 10000$; $j = 1, 2, \ldots$) follow a log-normal distribution where $Y^*_{rij} = \log Y_{rij}$ is, given covariates x_{1ri} and x_{2ri}, normal with mean

$$\mu^*_{ri} = \alpha^*_r + \beta_0 + \beta_1 x_{1ri} + \beta_2 x_{2ri} \tag{24}$$

and variance σ^2. The values α^*_r $(r = 1, \ldots, 10)$ in (24) were generated from a normal distribution with mean 0 and variance 0.084 and then centred about their mean, giving α^*_r = (- 0.366, -0.125, -0.119, -0.116, -0.055, -0.029, -0.006, 0.164, 0.318, 0.334) for $r = 1, \ldots, 10$. These values were chosen to represent a moderate amount (about 23%) of explained variation in the distribution of log duration times. The values for β_1, β_2 in (24) and for σ were chosen so that X_1, X_2 explained either 30% or 50% of the variation in log duration times, and β_0 was chosen to give a median jobless spell duration of 24 weeks. Choosing $\sigma^2_{y^*} = Var(Y^*_{rij}) = 0.36$ and noting that (a) $\sigma^2_{y^*} = Var(\alpha^*) + \beta_1^2 + \beta_2^2 + 2\rho\beta_1\beta_2 + \sigma^2$ and (b) the variation explained by X_1, X_2 is $EV = 1 - \sigma^2/\sigma^2_{y^*}$, we arrived at four scenarios given in Table 1. Durations are measured in weeks.

(ii) Durations for "not jobless" spells were, conditional on x_{1ri} and x_{2ri}, exponential with mean $\gamma_1 \exp(\gamma_2 x_{2ri})$, with $\gamma_1 = 11.619$, $\gamma_2 = 0.155$ (durations measured in weeks). This results in about 40% of the population experiencing at least one jobless spell over the six-year period.

(iii) Finally, LTF was generated for sampled individuals from a logistic regression model (7) in which

$$\text{logit}\,(\lambda_{rit}(\alpha)) = \alpha_{0t} + \alpha_{1t}\, x_{2ri} \quad t = 1, \ldots, 5, \tag{25}$$

where $\alpha_{0t} = 2.131$ and $\alpha_{1t} = -0.536$ $(t = 1, \ldots, 5)$. This results in about 50% of individuals LTF before year 6. We estimated the LTF probabilities $p_{it}(\alpha)$ assuming (25) to be true.

The finite population duration distribution in (2) is based on the jobless spells for the 100,000 population members under each scenario. The distribution $S_P(y)$ uses data extending beyond the 6 years of followup, for spells that were still ongoing at the 6 years. Because the duration distributions are stable over time, this does not disadvantage the estimators, which are based only on data over the 6 years.

Table 2 shows bias, standard deviation, average standard error and nominal 0.95 confidence interval coverage for the weighted KM estimators of (2) at selected durations y. In both scenarios I and IV the fact that X_2 affects both duration and LTF results in bias when only design weights are employed in (11) and (12). For the combined design-IPC weighting, the average of the standard errors based on (18) and (23) is slightly smaller than the standard deviation of $\hat{S}(y)$ across the 1000 samples in Scenario I. This produces slight under-coverage for the nominal 0.95 confidence intervals, which were based on treating $\log\{-\log \hat{S}(y)\}$ as approximately normal (Lawless 2003a, Sec. 3.2.3.1). There is also slight under-coverage at shorter durations across other scenarios. Confidence interval coverage for the design-weighted estimator is much less than 0.95, on the other hand, ranging from 0.70 to 0.80 for scenario I and from 0.82 to 0.92 for Scenario IV (results not shown in Table 2).

In Scenario II, with $\beta_2 = 0$ and $\rho = 0$, design weights on their own are satisfactory, but the inclusion of IPC weights does not reduce performance. For Scenario II we see that both design and combined design-IPC weighting performs well. In Scenario III with $\beta_2 = 0$ and $\rho = 0.3$, design weights alone are once again sufficient, but the simulation results show that the combined weights estimator performs equally well.

On a final note we remark that the use of IPC weights alone in these scenarios is not satisfactory, because of the variation in jobless spell duration distributions across the strata. Design weighting adjusts for this, as does the use of stratified variance estimation.

5 Application: Jobless Spell Durations from SLID

We apply the weighted KM methodology to the estimation of jobless spell durations for residents of Ontario and Quebec, based on the 1999 panel of Statistics Canada's Survey of Labour and Income Dynamics (SLID). Details concerning SLID are available from the Statistics Canada web site (www.statcan.gc.ca). Panel members were followed from 1999 to 2004 if not LTF earlier and for illustration, we consider jobless spells starting in 1999 and in 2000, for each province. It was necessary to drop some individuals for whom the start date of a spell was missing due to non-response. This left 359, 170, 311 and 211 spells for Ontario (1999, 2000) and Quebec (1999, 2000) respectively.

In SLID the top level strata are "economic units", of which there are 11 in Ontario and 17 in Quebec. The primary sampling units (PSUs) are geographic blocks known as dissemination areas and from each stratum, dissemination areas are chosen and then households are sampled within these areas. In variance estimation calculations we treated the economic units as the strata and PSUs as clusters. This allows for association of spell durations within individuals, households and PSUs. The 1999 panel (across all provinces) had 43,683 individuals in total, with about 28% LTF before year 6 (2004). Some individuals were missing at one annual interview but reappeared later; in our procedures such persons are treated as LTF from their first missed interview and the effective LTF rate is then about 42%.

Loss to followup was modelled separately for Ontario and Quebec using logistic regression models (7). Models included covariates based on sex, age, education level, marital and immigration status, household size and family composition. They were also based on whether the person was a student, a renter, resided in a urban area, and on whether the person was employed as of the previous year's interview. The models were based on a process of variable selection and diagnostic checks (Hosmer and Lemeshow 2000). Design weights for the weighted KM estimation were the 1999 SLID weights,

Table 2: Estimation of $S_P(y)$ for simulated jobless spell durations

		Bias		Std. Dev.	Aver. SE[1]	Cov[2]
y (weeks)	$S_P(y)$	DES	COMB	COMB	COMB	COMB
Scenario I						
9	0.914	-0.012	0.000	0.0137	0.0126	0.930
15	0.724	-0.025	0.002	0.0223	0.0207	0.932
21	0.523	-0.032	0.002	0.0248	0.0239	0.947
30	0.302	-0.029	0.002	0.0239	0.0228	0.932
47	0.106	-0.017	0.001	0.0174	0.0165	0.938
Scenario II						
10	0.904	0.001	0.001	0.016	0.015	0.924
16	0.717	0.001	0.002	0.024	0.022	0.939
22	0.524	0.001	0.002	0.026	0.025	0.931
31	0.310	0.003	0.004	0.024	0.023	0.939
49	0.104	0.002	0.001	0.016	0.015	0.945
Scenario III						
9	0.923	-0.003	0.000	0.014	0.013	0.928
16	0.706	-0.006	0.001	0.024	0.022	0.930
22	0.517	-0.007	0.002	0.025	0.025	0.948
31	0.303	-0.006	0.002	0.023	0.023	0.958
49	0.104	-0.003	0.001	0.015	0.015	0.949
Scenario IV						
9	0.920	-0.008	0.000	0.0136	0.0138	0.934
15	0.732	-0.018	0.000	0.0222	0.0212	0.937
22	0.503	-0.021	0.002	0.0246	0.0245	0.948
30	0.311	-0.019	0.002	0.0230	0.0232	0.956
48	0.104	-0.010	0.001	0.0160	0.0160	0.944

Results from 1,000 stratified random samples of size 1,450

[1] Average of estimated standard deviations for $\hat{S}(y)$ based on (18) and (23)

[2] Coverage for nominal 0.95 confidence interval for $S_P(y)$

which reflect the sampling design plus some calibration. Figure 2 (top panels) shows unweighted KM estimates and two weighted KM estimates for Ontario residents, based on (i) design weights only, and (ii) combined design-IPC weights. The bottom panels show pointwise standard errors for $\hat{S}(y)$ for (i) and (ii); for (ii) both the proper standard errors treating the IPC weights as estimates, and "naive" standard errors, treating them as known, are shown. The latter are given by (18) and (23) with the second term in the curly brackets in (18) set equal to zero. We see that the naive standard errors are very similar to those for the "design weights only" estimator, for which the weights are also treated as known. As (18) suggests, the proper standard errors for (ii) are smaller than the naive ones.

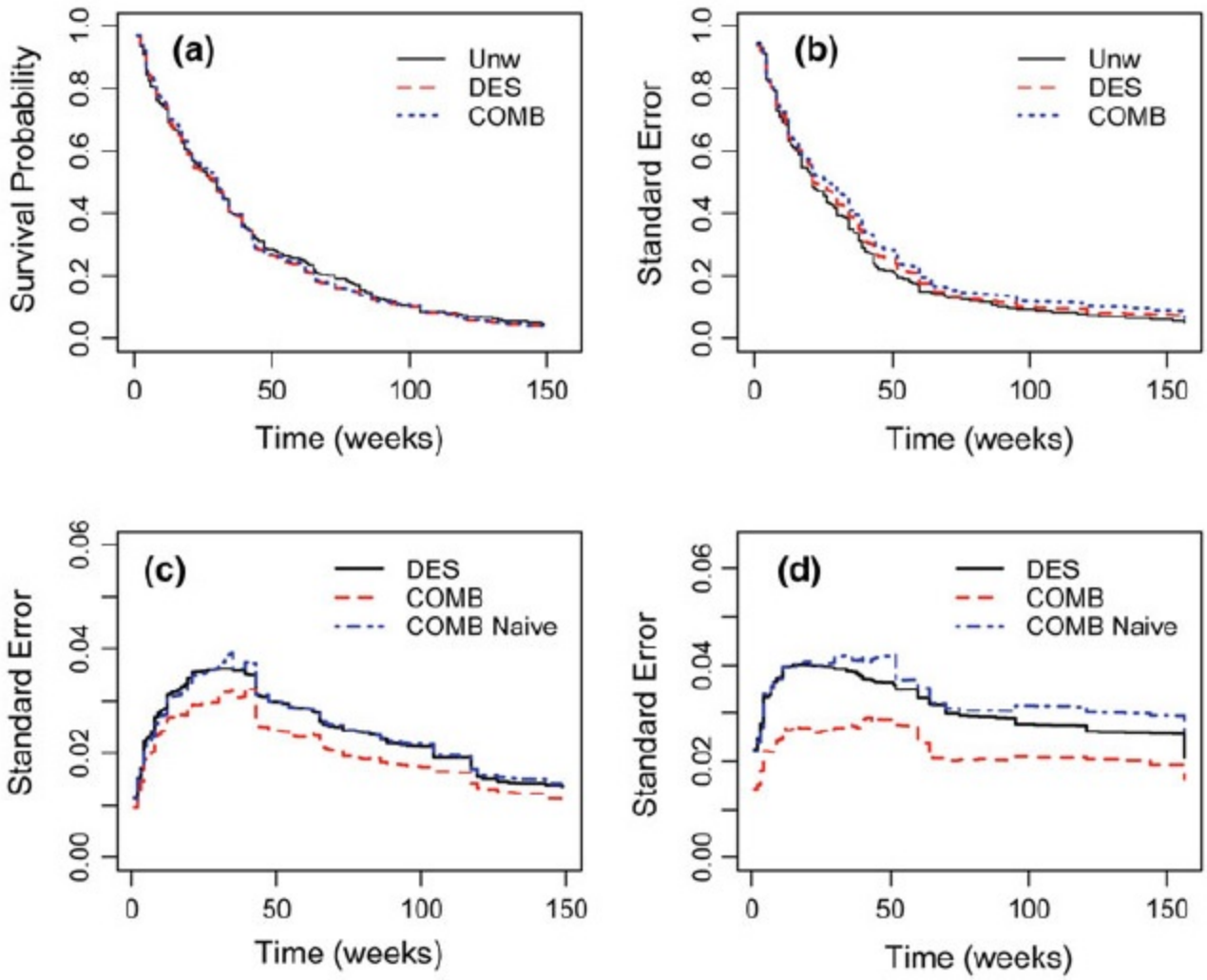

Figure 2: Weighted KM estimates for jobless spell duration probabilities and point-wise standard errors; (a), (c) for spells starting in 1999 (359 spells and 283 clusters); (b), (d) for spells starting in 2000 (270 spells and 220 clusters), respectively

Although the two weighted estimates in Figure 2 are similar, confidence intervals differ somewhat because of the differences in standard errors. A similar figure for Quebec residents shows slightly more difference in the estimates $\hat{S}(y)$. As a summary, Table 3 shows estimated median jobless duration and approximate 95% confidence intervals for each province and year. The confidence intervals were determined as the set of y-values satisfying -1.96 $\leq Z \leq$ 1.96, where $Z = \{\hat{S}(y) - 0.5\}/se\big(\hat{S}(y)\big)$ (Lawless 2003a, p. 93).

Figure 1 and Table 3 indicate that jobless spells starting in 2000 tend to be shorter than those starting in 1999, for each province. Some of this could be due to persons jobless in a given year being more likely LTF at the next interview. This type of LTF is "not missing at random" (NMAR) and is not handled by the IPC weights, which adjust only for the employment status at the preceding interview. We comment further on this at the end of Section 6. A plot like Figure 1 that was based on separate random samples of jobless spells starting in (a) 1999 and (b) 2000, and with perfect followup, would not necessarily show that spells in 2000 tended to be shorter.

Table 3: Estimated median jobless duration, Ontario and Quebec

		Ontario		Quebec	
Year	Method	Estimate[1]	CI	Estimate	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)

[1] Durations are in weeks

6 CONCLUDING REMARKS

This paper provides weighted KM estimates for finite-population survivor functions of spell durations. In many contexts an individual who is in a particular state can make a transition to any of several other states; for example, a jobless individual may obtain a job, become self-employed or leave the labor force. In this case we may, if we wish, consider the duration Y in the competing risks sense, with a variable L denoting which new state results from the transition. Suppose there are L_0 other states that can be entered from the current state and denote

$$h_\ell(y) = \Pr\left(Y = y, L = \ell | Y \geq y\right) \qquad \ell = 1, \ldots, L_0 \tag{26}$$

as "cause-specific" hazard functions corresponding to (4). The sub-probability distribution for duration Y when state ℓ is entered next is then (Lawless 2003a, Chap. 9)

$$f_\ell(y) = \Pr\left(Y = y, L = \ell\right) = h_\ell(y) S(y), \qquad \ell = 1, \ldots, L_0 \tag{27}$$

where $S(y)$ is the marginal survivor function in (3) without regard to the next state occupied. The sub-distribution functions are

$$F_\ell(y) = \Pr\left(Y \leq y, L = \ell\right) = \sum_{s=1}^{y} f_\ell(s), \qquad \ell = 1, \ldots, L_0. \tag{28}$$

The discrete hazard functions $h_\ell(y)$ are estimated exactly as in Section 3, except with $d_{ijt}(y)$ in (9) and later expressions replaced by

$$d^\ell_{ijt}(y) = I\left(Y_{ij} = y, L_{ij} = \ell, t - 1 < u_{ij} + y \leq t\right) R_{it}, \tag{29}$$

where L_{ij} is the state occupied following the jth occupation of the base state (e.g. jobless) for individual i. Note that in estimating $f_\ell(y)$ in (27), $\sum_{\ell=1}^{L_0} \hat{h}_\ell(y) = \hat{h}(y)$ in (11) and that (12) gives $\hat{S}(y)$.

Variance estimation is slightly more involved in this case since the vector θ consists of $L_0 T$ components $h_\ell(y)$ ($\ell = 1, \ldots, L_0$; $y = 1, \ldots, T$). Some simplification along lines indicated in Appendix 2 can be made if we use the fact that the matrix $A_{11}(\theta, \alpha)$ is block diagonal, with the blocks corresponding to the values $\ell = 1, \ldots, L_0$. Variance estimation for $\hat{F}_\ell(y)$, given by (28) with estimates for the $h_\ell(y)$ and $S(y)$ inserted, follows from the delta theorem, since

$$\hat{F}_\ell(y) = \sum_{s=1}^{y} \hat{h}_\ell(s) \prod_{u=1}^{s-1} \left\{ 1 - \sum_{\ell'=1}^{L_0} \hat{h}_{\ell'}(u) \right\}. \tag{30}$$

Other approaches to estimation of a duration distribution from survey data can be considered. One is to use a regression model $S(y|x_i)$, which might possibly improve precision while adjusting for some design factors and possibly dropout (e.g. Korn and Graubard 1999). For example, if each individual in a population of size N has exactly one duration Y_i, then $S(y)$ could be estimated as $\sum_{i=1}^{n} \pi_i^{-1} \hat{S}(y|x_i)/N$. It would generally be necessary, however, to adjust for time-varying factors related to loss to followup (Hajducek and Lawless, 2012) in estimating $S(y|x)$. In addition, it is not obvious how to deal with settings where only some individuals experience a spell, and this approach is less easily adaptable to settings where individuals can have multiple spells, where it might be unappealing to model different spells with a single regression model. Regression estimation of finite population quantities (e.g. Lumley et al. 2011) has not been applied to such settings.

Similarly, attempts to develop augmented estimators (e.g. Van der Laan et al. 2002; Rotnitzky 2009; Stitelman and van der Laan 2010) could be undertaken. This would involve combining weighted KM estimation and regression estimation. This is feasible if an individual can have exactly one spell. Then, for example, we could consider the estimator

$$\hat{S}(y) = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i}{\pi_i} \left\{ \frac{I(Y_i \geq y) I(\tilde{C}_i \geq y)}{\tilde{G}_i(y)} + \frac{1 - I(\tilde{C}_i \geq y)}{\tilde{G}_i(y)} \hat{S}(y|x_i) \right\}, \tag{31}$$

where $\tilde{C}_i = C_i - U_i$ is the effective censoring time for Y_i (corresponding to a spell starting at time U_i) and $G_i(y) = Pr(\tilde{C} \geq y | Z_i^c)$ is a model for $\tilde{C}_i$ conditional on covariate histories $\{Z_i^c(t), t > 0\}$, as in Section 2. Once again, however, it is unclear how this approach can be applied when only some individuals have spells, and when there can be multiple spells per individual. In addition, one would have to deal with the sample design probabilities. Similar remarks apply to "enhanced" method of estimation such as those in Stitelman and van der Laan (2010). In that paper the authors provide a number of ways to estimate treatment specific survivor functions, but it is not clear how one could deal with the specific aspects of the problem here. The weighted KM estimates we provide are simple to implement and very useful for the types of applications considered here, and it is beyond our present scope to consider whether more complicated approaches could be modified or adapted to this setting. However, future investigations in this direction might be useful.

In many longitudinal surveys, including SLID, there may be missing information on certain variables for an individual in a given year, due to non-response or insufficient knowledge for a proxy respondent. In the case of jobless spells, start times U_{ij} are sometimes missing. This poses a difficult challenge which has not been addressed here or elsewhere in the literature. Imputation methods or likelihood methods implemented using EM algorithms require a joint model for the durations of spells in all states, thus involving much more detailed modeling. Another practical concern is measurement error in reported start

and end times for spells. For example, with widely spaced interviews a "seam" effect may occur, in which individuals tend to locate events closer to interview times than they actually are (Callegaro 2008). Missing and mismeasured variables are data quality issues which are difficult to address methodologically, and every effort should bemade to design data collection so as to minimize such problems. Pyy-Martikainen and Rendtel (2008) provide a lucid discussion of measurement error issues and an empirical study based on a comparison of survey data and Finnish administrative register data on unemployment spells. They make a number of important suggestions concerning data collection.

It has been assumed that LTF at time t depends only on variables measured up to time $t-1$. This is necessary for the MAR assumption and application of IPCW methods, but when interviews are widely spaced there is likely to be some dependence on events and other variables for the time interval $(t-1, t]$. In some cases it may be possible to trace a random sample of persons LTF but, failing this, one can do sensitivity analysis (Scharfstein and Robins, 2002) or run simulations to study the effect of non-MAR factors. Another very important possibility is the use of administrative data to assess (and adjust) survey data. Pyy-Martikainen and Rendtel (2009) carry out an empirical study of the effects of both initial nonresponse and LTF, using the survey and administrative data mentioned above. They find that in their setting, initial nonresponse and LTF are both significant sources of bias.

Finally, this paper deals with estimation of a finite population duration distribution. If we were instead interested in the durations of successive spells (and the occurrence of such spells) for individuals, then a different approach involving models for the multi-state processes generating the spells would be needed. This is more complex; see Hajducek and Lawless (2012) and Pyy-Martikainen and Rendtel (2008) for some comments and references. If we are interested in marginal duration distribution for spells in a given state, then methods such as those in Satten and Datta (2002) may be useful.

ACKNOWLEDGEMENTS

We are grateful to Mary Thompson and Christian Boudreau for valuable suggestions and comments. We also thank Georgia Roberts for her guidance with regard to data from SLID. Sincere thanks are extended to Pat Newcombe-Welch from the South-Western Ontario Research Data Centre (SWORDC) for her timely assistance.

This research was supported by the National Council for Science and Technology of Mexico (CONACyT), the MITACS Network of Centres of Excellence, and the Natural Sciences and Engineering Research Councilof Canada. We thank Statistics Canada for hosting an internship for DMH, and the South-Western Ontario Research Data Centre (SWORDC) for providing access to SLID data.

A APPENDIX

A.1 UNBIASEDNESS OF ESTIMATING EQUATIONS (10)

Consider a random member of the population P, and let $H_i(M)$ denote the full duration history $\{D_i(1), \ldots, D_i(M)\}$ for individual i over $[0, M]$. Using (9), we rewrite (8) as

$$U_i^W(y) = \sum_{t=1}^{M} \sum_{j=1}^{m_i} \frac{\delta_{ijt}(y)}{\pi_i p_{it}(\alpha)} \left[d_{ijt}(y) - h(y) \right], \quad y = 1, 2, ...T \tag{A1}$$

and evaluate the expectation of the tth term of $R_i U_i^W(y)$ as

$$E_{H_i(M),Z_i^c(t),Z_i^D}\ E\left\{R_i \sum_{j=1}^{m_i} \frac{\delta_{ijt}(y)}{\pi_i p_{it}(\alpha)} \left[d_{ijt}(y) - h(y)\right] \Big| Z_i^c(t), Z_i^D\right\},$$

where $R_i = I(\text{person } i \text{ is in the panel sample})$ and Z_i^D is the set of design factors (stratum information) that specifies $\pi_i = \Pr(R_i = 1 | Z_i^D)$. It is assumed that R_i is independent of $H_i(M)$ and $Z_i^c(t)$, given Z_i^D and that $R_{it} = I(C_i \geq t)$ is independent of $H_i(M)$, given $Z_i^C(t)$ and $R_i = 1$. The latter is the MAR assumption of Section 3. In addition, we consider $H_i(M)$ as random in the superpopulation framework, but comment below on the finite population case. The above expectation then equals

$$E_{H_i(M)}\left\{\sum_{j=1}^{m_i} I\left(Y_{ij} \geq y, t-1 < u_{ij} + y \leq t\right)\left[d_{ijt}(y) - h(y)\right]\right\},$$

and the expectation of (A1) is

$$\begin{aligned} & E_{H_i(M)}\left\{\sum_{j=1}^{m_i}\sum_{t=1}^{M} I\left(Y_{ij} \geq y\right) I\left(t-1 < u_{ij} + y \leq t\right)\left[d_{ijt}(y) - h(y)\right]\right\} \\ = \quad & E_{H_i(M)}\left\{\sum_{j=1}^{m_i} I\left(Y_{ij} \geq y\right)\left[d_{ij}(y) - h(y)\right]\right\} \end{aligned}$$

where $d_{ij}(y) = I(Y_{ij} = y)$. Note that $m_i \geq 0$ and the Y_{ij} (for $j = 1, \ldots, m_i$) are random variables. The expectation above is, according to the definitions for $S_P(y)$ and $S(y)$ in Section 2,

$$E_{H_i(M)}\left\{\sum_{j=1}^{m_i} I\left(Y_{ij} = y\right) - h(y) I\left(Y_{ij} \geq y\right)\right\} = E_{H_i(M)}\left\{m_i f(y) - m_i S(y) h(y)\right\} = 0.$$

More directly, in terms of the finite population $S_P(y)$ the $H_i(M)$ $(i = 1, 2, ..., N^*)$ are fixed finite population quantities, and the expectation of $\sum_{i=1}^{N^*} R_i U_i^W(\theta)$ has terms

$$\sum_{i=1}^{N^*}\sum_{j=1}^{m_i}\left\{I(Y_{ij} = y) - I(Y_{ij} \geq y) h_P(y)\right\}$$

which equals zero by the definition of $S_P(y)$ and $h_P(y)$. Thus, (8) has expectation 0 with respect to the sampling plan giving R_i and random LTF process.

A.2 Variance estimates for $\hat{\theta}$

Variance estimates for $\hat{\theta}$ follow directly from asymptotic theory for estimating functions (White 1982), as we indicate below. For the case of longitudinal data with LTF, Robins et al. (1995) obtained an estimate from first principles. Adapted to our survey sampling and duration distribution context, this results in (18). Robins et al. and others (e.g. Miller et al. 2001) express it a little differently, with the center term in (18) given as (in our case)

$$\sum_{r=1}^{R} \frac{K_r}{K_r - 1} \sum_{k=1}^{K_r}\left[E_{rk.}\left(\hat{\theta}, \hat{\alpha}\right) - \bar{E}_r\left(\hat{\theta}, \hat{\alpha}\right)\right]\left[E_{rk.}\left(\hat{\theta}, \hat{\alpha}\right) - \bar{E}_r\left(\hat{\theta}, \hat{\alpha}\right)\right]',$$

where $E_{rk.}\left(\hat{\theta},\hat{\alpha}\right)$ and $\bar{E}_r\left(\hat{\theta},\hat{\alpha}\right)$ come from terms

$$E_{rki}\left(\theta,\alpha\right)=U_{rki}\left(\theta,\alpha\right)-B_{12}\left(\theta,\alpha\right)B_{22}\left(\alpha\right)^{-1}G_{rki}\left(\alpha\right). \tag{A2}$$

This results in a slightly different estimate than (18).

Alternatives to (18) also come from a direct application of the results of White (1982) to the estimating functions in (13) and (14). The asymptotic covariance matrix for $\hat{\psi}=\left(\hat{\theta}',\hat{\alpha}'\right)'$ is consistently estimated by

$$\widehat{Var}\left(\hat{\psi}\right)=A\left(\hat{\psi}\right)^{-1}B\left(\hat{\psi}\right)^{-1}A\left(\hat{\psi}\right)^{-1}, \tag{A3}$$

where, in partitioned form,

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

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

Using the fact that

$$A(\psi)^{-1}=\left(\begin{array}{cc}A_{11}\left(\theta,\alpha\right)^{-1} & -A_{11}\left(\theta,\alpha\right)^{-1}A_{12}\left(\theta,\alpha\right)A_{22}(\alpha)^{-1} \\ 0 & A_{22}(\alpha)^{-1}\end{array}\right)$$

we obtain $\widehat{Var}(\hat{\theta})$ as the upper left block of (A2) evaluated at $\left(\hat{\theta},\hat{\alpha}\right)$:

$$\widehat{Var}(\hat{\theta})=A_{11}\left(\hat{\theta},\hat{\alpha}\right)^{-1}\left\{B_{11}\left(\hat{\theta},\hat{\alpha}\right)-A_{12}\left(\hat{\theta},\hat{\alpha}\right)A_{22}(\hat{\alpha})^{-1}B_{21}\left(\hat{\theta},\hat{\alpha}\right)\right\}A_{11}\left(\hat{\theta},\hat{\alpha}\right)^{-1}, \tag{A4}$$

where $B(\hat{\psi})$ is an estimate of $B(\psi)$. The estimate (A4) differs from (18) in having $A_{12}(\hat{\theta},\hat{\alpha})$ and $A_{22}(\hat{\alpha})$ in place of $B_{12}(\hat{\theta},\hat{\alpha})$ and $B_{22}(\hat{\alpha})$, respectively. The estimates are asymptotically equivalent. This is shown by noting that (i) $G(\alpha)$ is a sum of likelihood estimating functions, and so it follows that $E\{A_{22}(\alpha)\}=\{B_{22}(\alpha)\}$, and (ii) in the terms of $G(\theta,\alpha)$ in (13), the conditional independence of the random variables R_{rkijt} and $\{d_{rkijt}(y),I(Y_{rkij}\geq y,t-1<u_{rkij}+y\leq t)\}$ given $Z^c_{rki}(t)$ implies that $E\{A_{12}(\theta,\alpha)\}=E\{B_{12}(\theta,\alpha)\}$. In particular regarding (ii), we note that (replacing rki with i for convenience) the ith terms of $A_{12}(\theta,\alpha)$ are

$$\left(A_{12i}\right)_{ys}=-\frac{\partial U_i(\theta,\alpha)_y}{\partial\alpha_s}=I(m_i>0)\sum_{t=1}^{M}\sum_{j=1}^{m_i}w_{ijt}(y)\frac{\partial\log p_{it}(\alpha)}{\partial\alpha_s}\left[d_{ijt}(y)-h(y)\right]$$

for $y=1,\ldots,T$ and $s=1,\ldots,M$. Now,

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

under the logistic model (7). Thus

$$(A_{12i})_{y,s} = I(m_i > 0) \sum_{j=1}^{m_i} \sum_{t=1}^{M} I(s \leq t) w_{ijt}(y) Z_i^c(s) \left[1 - \lambda_{is}(\alpha_s)\right] \left[d_{ijt}(y) - h(y)\right] \tag{A5}$$

In addition, the ith terms in $B_{12}(\theta, \alpha)$ are

$$\begin{aligned} (B_{12i})_{y,s} &= U_i(\theta, \alpha)_y G_i(\alpha)_s \\ &= I(m_i > 0) \sum_{j=1}^{m_i} \sum_{t=1}^{M} w_{ijt}(y) R_{i,s-1} Z_i^c(s) \left\{R_{is} - \lambda_{is}(\alpha_s)\right\} \left\{d_{ijt}(y) - h(y)\right\}. \end{aligned}$$

For $s > t$, the tth term in $E\{(B_{12i})_{y,s}\} = 0$ since R_{is} is independent of the entire duration history $H_i(M)$, conditional on $Z_i^c(s)$. For $s \leq t$, we have $E\{R_{is} - \lambda_{is}(\alpha_s) | R_{i,s-1} = 1, R_{it} = 1\} = 1 - \lambda_{is}(\alpha_s)$, on the other hand. Thus,

$$\begin{aligned} E\left\{(B_{12i})_{y,s}\right\} &= E\left\{ I(m_i > 0) \sum_{j=1}^{m_i} \sum_{t=1}^{M} I(s \leq t) w_{ijt}(y) Z_i^c(s) \left[1 - \lambda_{is}(\alpha_s)\right] \left[d_{ijt}(y) - h(y)\right] \right\} \\ &= E\left\{(A_{12i})_{y,s}\right\}. \end{aligned}$$

Finally, we note that $A_{22}(\alpha)$ is a block diagonal matrix with blocks (see (16))

$$A_{22}^{(t)}(\alpha_t) = -\partial G^{(t)}(\alpha_t) / \partial(\alpha_t') \qquad t = 1, \ldots, M$$

of dimensions q_t $(t = 1, \ldots, M)$. Correspondingly, we can use the fact that the estimating functions $G^{(t)}(\alpha_t)$ are mutually independent for $t = 1, \ldots, M$ under the assumptions in the paper, to replace $B_{22}(\hat{\alpha})$ in (18) with a block diagonal version.

REFERENCES

Callegaro, M. (2008). Seam effects in longitudinal surveys. *Journal of Official Statistics* 24, 387–409

Folsom, R. and LaVange, L. and Williams, R.L. (1981). *A probability sampling perspective on panel data analysis.* In Panel Surveys, editors, D. Kasprzyk, G.J. Duncan, G. Kalton and M.P. Singh. 108–138, Wiley, New York.

Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression*, second edition, Wiley, New York.

Kalton, G. and Miller, D.P. and Lepkowski, J. (1992). Analyzing spells of program participation in the SIPP. Technical report, Survey Research Center, University of Michigan.

Korn, E.L. and Graubard, B.I. (1999). *Analysis of Health Surveys*, Wiley, New York.

Kovacevic, M.S. and Roberts, G. (2007). Modelling durations of multiple spells from longitudinal survey data, *Survey Methodology*, 33,13–22.

Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data.* Wiley, New York, second edition.

Lawless, J.F. (2003). Censoring and Weighting in Survival Estimation from Survey Data, In *Proceedings of the Survey Methods Section*, Statistical Society of Canada 2003 Annual Meeting.

Miller, M. E., Ten Have T.R., Reboussin, B.A., Lohman, K.K. and Rejeski, W. J. (2001). A marginal model for analyzing discrete outcomes from longitudinal surveys with outcomes subject to multiple cause nonnesponse. *Journal of the American Statistical Association*, 96, 844–857.

Plewis, I. (2007). Non-Response in a Birth Cohort Study: The Case of the Millennium Cohort Study. *International Journal of Social Research Methodology*, 10, 325–334.

Pyy-Martikainen, M. and Rendtel, U. (2008). Measurement errors in retrospective reports of event histories. A validation study with Finnish register data. *Survey Research Methods*, 3, 139–155.

Pyy-Martikainen, M. and Rendtel, U. (2009). Assessing the impact of initial nonresponse and attrition in the analysis of unemployment duration with panel surveys. *Advances in Statistical Analysis*, 92, 293–318.

Robins, J.M. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the Biopharmaceutical Section*, 24–33. American Statistical Association, Alexandria VA.

Robins, J.M. and Finkelstein D.M. (2000). Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*, 56, 779–788.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106–121.

Rotnitzky, A. (2009). Inverse probability weighted methods. In Fitzmaurice,G., Davidian, M., Verberke, G. & Molenberghs, G. (Ed.) *Longitudinal Data Analysis*, Chapman and Hall/CRC, Boca Raton.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581–592.

Satten, G.A. and Datta, S. and Robins, J. (2001). Estimating the marginal survival function in the presence of time dependent covariates. *Statistics and Probability Letters*, 54, 397–403.

Scharfstein, D.O. and Robins, J.M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, 89, 617–634.

Van der Laan, M.J., Hubbard, A. and Robins, J. (2002). Locally efficient estimation of a multivariate survivor function in longitudinal studies. *Journal of the American Statistical Association*, 99, 494–507.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.

Williams, R.L. (1995). Product -limit survival functions with correlated survival times. *Lifetime Data Analysis*, 1, 171–186.