Survey Methodology, June 2007
Vol. 33, No. 1, pp. 13-22
Statistics Canada, Catalogue No. 12-001

Modelling durations of multiple spells from longitudinal survey data

Milorad S. Kovačević and Georgia Roberts [1]

Abstract

We investigate some modifications of the classical single-spell Cox model in order to handle multiple spells from the same individual when the data are collected in a longitudinal survey based on a complex sample design. One modification is the use of a design-based approach for the estimation of the model coefficients and their variances; in the variance estimation each individual is treated as a cluster of spells, bringing an extra stage of clustering into the survey design. Other modifications to the model allow a flexible specification of the baseline hazard to account for possibly differential dependence of hazard on the order and duration of successive spells, and also allow for differential effects of the covariates on the spells of different orders. These approaches are illustrated using data from the Canadian Survey of Labour and Income Dynamics (SLID).

Key Words: Cox regression; Design-based inference; Model-based inference; Spell order; SLID.

1. Introduction

The modelling problem addressed in this paper is known under different names such as correlated failure-time modelling, multivariate survival modelling, multiple spells modelling, or a recurrent events problem. It has been studied in the biomedical (*e.g.*, Lin 1994, Hougaard 1999), social (Blossfeld and Hamerle 1989, Hamerle 1989) and economic literature (Lancaster 1979, Heckman and Singer 1982). Generally this type of modelling is required to address issues that arise in time-to-event studies when two or more events occur to the same subject and where the research interest is to assess the effect of various covariates on the length of a spell. Failure times are correlated within a subject, and thus the assumption of independence of failure times conditional on given measured covariates, required by standard survival models, is likely to be violated. In studies of duration of spells (poverty, unemployment, *etc.*), the "failure" is equivalent to exiting from the state of interest. An additional property of many multiple spells, often ignored, is that the spells are ordered "events"; that is, the second spell cannot occur before the first, *etc*. This paper was motivated by a study of unemployment spells, discussed further in Section 5.

The dependence among the spells from the same individual arises from the fact that these spells share certain unobserved characteristics of the individual. The effect of these unobserved characteristics can be explicitly modelled as a random effect (*e.g.*, Clayton and Cuzick 1985). When this is done, it is assumed that the random effect follows a known statistical distribution. The gamma distribution with mean 1 and unknown variance is the distribution of choice in many applications. Then, estimates of random and fixed effects can be obtained by some suitable method (*e.g.*, two-stage likelihood (Lancaster 1979), using an EM algorithm (Klein 1992), *etc.*). This paper does not explore this approach.

Another approach that has been taken - and is the one that we will be using - is to take a semi-parametric approach where we do not explicitly model the dependence among multiple spells. We model the marginal distributions of the individual spells, with a possible utilization of the order of the spells in the model specification. In the non-survey context, Lin (1994) describes how it is sufficient just to modify the "naïve" covariance matrix of the estimated model coefficients obtained under the assumption of independence since the correlated durations need to be accounted for in the variance estimates but not in the estimates of coefficients per se.

In socio-economic studies of spell durations the data sources are frequently longitudinal surveys with complex sample designs that involve stratification, sampling in several stages, selection with unequal probabilities, stochastic adjustments for attrition and non-response, calibration to known parameters, *etc*. Consequently, it is necessary to account for the impact of the sample design on the distribution of the sample data when estimating model parameters and the variances of these estimates. Our approach when analyzing complex survey data is to model the marginal distributions of the multiple spells using single-spell methods, treating the dependence among the spells as a nuisance - both the dependence due to the correlation of spells from the same person and dependence among individuals due to the survey design - but taking account of the unequal selection probabilities through the survey weights. Based on the model chosen, finite population

1. Milorad S. Kovačević, Methodology Research Advisor, Statistics Canada, Ottawa, Canada, K1A 0T6. E-mail: kovamil@statcan.ca; Georgia Roberts, Chief of the Data Analysis Resource Center at the Social Survey Methods Division, Statistics Canada, Ottawa, Canada, K1A 0T6. E-mail: robertg@statcan.ca.

parameters are defined and estimated as in Binder (1992). Standard errors are estimated using an appropriate design-consistent linearization method under the assumption that the primary sampling units are sampled with replacement within strata. This assumption is viable when the sampling rates at the first stage are small, as is generally the case in socio-economic surveys. Also, for such samples, the difference between finite population and superpopulation inference (*i.e.*, the standard errors and the test statistics) has been found to be rather negligible (Lin 2000). Therefore, the results from inference based on our approach extend beyond the finite population under study.

In the next section we review single-spell modelling and some methods for robust estimation of variances when the model is misspecified - first under a model-based framework and then under a design-based one. Section 3 contains further discussion of robust variance estimation for multiple spells. In Section 4, we introduce three models for multiple spells and describe how to fit these models using design-based robust estimation methods. In Section 5, we fit these models to data from the Canadian Survey of Labour and Income Dynamics (SLID) and discuss the results. Finally, Section 6 contains some overall remarks.

2. Inference for the single-spell hazard rate model

The duration of a spell (or simply, a spell) experienced by an individual is a random variable denoted by T. We are particularly interested in the hazard function $h(t)$ of T at time t, defined as the instantaneous rate of spell completion at time t given that it has not been completed prior to time t, formally

$$h(t) = \lim_{dt \to 0} \frac{\text{Prob}\{t \le T < t + dt \mid T \ge t\}}{dt}.$$

The value of the hazard function at t is called the exit rate to emphasize that the completion of the spell is equivalent to exiting the state of interest. Duration models and analysis of duration in general are formulated and discussed in terms of the hazard function and its properties.

From a subject matter perspective, frequently the main interest is to study the impact of some key covariates on the distribution of T. A proportional hazards model is often chosen for such a study. Under the proportional hazards model, the hazard function of the spell T given a vector of possibly time-varying covariates $\mathbf{x}(t) = (x_1(t), ..., x_p(t))'$ is

$$h(t \mid \mathbf{x}(t)) = \lambda_0(t)\, e^{\mathbf{x}'(t)\boldsymbol{\beta}}. \tag{1}$$

The function $\lambda_0(t)$ is an unspecified baseline hazard function and gives the shape of $h(t \mid \mathbf{x}(t))$. The baseline hazard describes the duration dependence, such as whether the hazard rate depends on time already spent in the spell. For example, negative dependence describes the situation where the longer the spell the smaller the probability of exit. If an individual has all $\mathbf{x}(t)$ variables set at 0, the value (level) of the hazard function is equal to the baseline hazard.

2.1 Model-based inference

The vector $\boldsymbol{\beta}$ contains the unknown regression parameters showing the dependence of the hazard on the $\mathbf{x}(t)$ vector, and may be estimated by maximizing the partial likelihood function (Cox 1975):

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{e^{\mathbf{x}_i'(T_i)\boldsymbol{\beta}}}{\sum_{j=1}^{n} Y_j(T_i)\, e^{\mathbf{x}_j'(T_i)\boldsymbol{\beta}}} \right]^{\delta_i}. \tag{2}$$

Here $T_1, ..., T_n$ are n possibly right-censored durations; $\delta_i = 1$ if T_i is an observed duration and $\delta_i = 0$ otherwise; and $\mathbf{x}_i(t)$ is the corresponding covariate vector observed on $[0, T_i]$. The denominator sum is taken over the spells that are at risk of being completed at time T_i, *i.e.*, $Y_j = 1$ if $t \le T_j$, and is equal to 0 otherwise. The estimate $\hat{\boldsymbol{\beta}}$ of the model parameter $\boldsymbol{\beta}$ is obtained by solving the partial likelihood score equation

$$U_0(\boldsymbol{\beta}) = \sum_{i=1}^{n} u_{i0}(T_i, \boldsymbol{\beta}) = 0, \tag{3}$$

where

$$u_{i0}(T_i, \boldsymbol{\beta}) = \delta_i \left\{ \mathbf{x}_i(T_i) - \frac{S^{(1)}(T_i, \boldsymbol{\beta})}{S^{(0)}(T_i, \boldsymbol{\beta})} \right\}, \tag{4}$$

$$S^{(0)}(t, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t)\, e^{\mathbf{x}_i'(t)\boldsymbol{\beta}}, \tag{5}$$

and

$$S^{(1)}(t, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t)\, \mathbf{x}_i(t)\, e^{\mathbf{x}_i'(t)\boldsymbol{\beta}}. \tag{6}$$

If model (1) is true and the durations are independent, the model-based variance matrix of the score function $U_0(\boldsymbol{\beta})$ is

$$J(\boldsymbol{\beta}) = -\partial U_0(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$$
$$= \sum_{i=1}^{n} \delta_i \left\{ \frac{S^{(2)}(T_i, \boldsymbol{\beta})}{S^{(0)}(T_i, \boldsymbol{\beta})} - \frac{S^{(1)}(T_i, \boldsymbol{\beta})[S^{(1)}(T_i, \boldsymbol{\beta})]'}{[S^{(0)}(T_i, \boldsymbol{\beta})]^2} \right\},$$

where

$$S^{(2)}(t, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t)\, \mathbf{x}_i(t)\, \mathbf{x}_i'(t)\, e^{\mathbf{x}_i'(t)\boldsymbol{\beta}}.$$

The approximate variance of $\hat{\boldsymbol{\beta}}$, obtained by linearization, is $J^{-1}(\hat{\boldsymbol{\beta}})$.

If the form of (1) is incorrect but observations are independent, Lin and Wei (1989) provide the robust variance estimator for $\hat{\boldsymbol{\beta}}$ as

$$J^{-1}(\hat{\boldsymbol{\beta}})\, G(\hat{\boldsymbol{\beta}})\, J^{-1}(\hat{\boldsymbol{\beta}}), \tag{7}$$

where

$$G(\boldsymbol{\beta}) = \sum_{i=1}^{n} g_i(\boldsymbol{\beta})\, g_i'(\boldsymbol{\beta})$$

and

$$g_i(\boldsymbol{\beta}) = u_{i0}(T_i, \boldsymbol{\beta}) - \sum_{j=1}^{n} \delta_j \frac{Y_i(T_j)\, e^{\mathbf{x}_i'(T_j)\boldsymbol{\beta}}}{n S^{(0)}(T_j, \boldsymbol{\beta})} \left\{ \mathbf{x}_i(T_j) - \frac{S^{(1)}(T_j, \boldsymbol{\beta})}{S^{(0)}(T_j, \boldsymbol{\beta})} \right\}. \tag{8}$$

2.2 Design-based inference

For observations from a survey with a complex sample design, Binder (1992) used a pseudo-likelihood method to estimate the parameters and their variances for a proportional hazards model in the case of a single spell per individual. In particular, he first defined the finite population parameter of interest as a solution of the partial likelihood score equation (3) calculated from the spells of the finite population targeted by the survey:

$$U_0(\mathbf{B}) = \sum_{i=1}^{N} u_{i0}(T_i, \mathbf{B}) = 0,$$

where $u_{i0}(T_i, \mathbf{B})$ is the score residual defined in the same way as $u_{i0}(T_i, \boldsymbol{\beta})$, except that the averages in the definitions of $S^{(0)}(t, \mathbf{B})$ and $S^{(1)}(t, \mathbf{B})$ extend over N observations rather than n. Note that if all members of the finite population targeted by the survey do not experience spells, N represents the size of the subpopulation that experiences spells, and the summation is over these N individuals.

An estimate $\hat{\mathbf{B}}$ of the parameter $\mathbf{B}$ is obtained as a solution to the partial pseudo-score estimating equation

$$\hat{U}_0(\hat{\mathbf{B}}) = \sum_{i=1}^{N} w_i(s)\, \hat{u}_{i0}(T_i, \hat{\mathbf{B}}) = 0,$$

where $w_i(s) = w_i$, the survey weight, if $i \in s$, and 0 otherwise. Function $\hat{u}_{i0}(T_i, \hat{\mathbf{B}})$ takes the form

$$\hat{u}_{i0}(T_i, \hat{\mathbf{B}}) = \delta_i \left\{ \mathbf{x}_i(T_i) - \frac{\hat{S}^{(1)}(T_i, \hat{\mathbf{B}})}{\hat{S}^{(0)}(T_i, \hat{\mathbf{B}})} \right\},$$

where

$$\hat{S}^{(0)}(t, \hat{\mathbf{B}}) = \sum_{i=1}^{N} w_i(s)\, Y_i(t)\, e^{\mathbf{x}_i'(t)\hat{\mathbf{B}}},$$

and

$$\hat{S}^{(1)}(t, \hat{\mathbf{B}}) = \sum_{i=1}^{N} w_i(s)\, Y_i(t)\, \mathbf{x}_i(t)\, e^{\mathbf{x}_i'(t)\hat{\mathbf{B}}}.$$

Generally, the design-based variance of an estimate $\hat{\boldsymbol{\theta}}$ that satisfies an estimating equation of the form $\hat{U}(\hat{\boldsymbol{\theta}}) = \sum w_i\, u_i(\hat{\boldsymbol{\theta}}) = 0$ can be estimated, using linearization, as

$$\hat{J}^{-1}\, \hat{V}(\hat{U}(\hat{\boldsymbol{\theta}}))\, \hat{J}^{-1}, \tag{9}$$

where $\hat{J} = \partial \hat{U}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ is evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and $\hat{V}(\hat{U}(\hat{\boldsymbol{\theta}}))$ is the estimated variance of the estimated total $\hat{U}(\boldsymbol{\theta})$ obtained by some standard design-based variance estimation method (see for example Cochran (1977)) and evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Binder (1983) states that in order to use this approach to derive a consistent estimate of the variance, $\hat{U}(\hat{\boldsymbol{\theta}})$ must be expressed as a sum of independent random vectors. In the case of the proportional hazards model above, $\hat{U}_0(\hat{\mathbf{B}})$ does not satisfy this condition since each $\hat{u}_{i0}$ is a function of $\hat{S}^{(0)}(T_i, \hat{\mathbf{B}})$ and $\hat{S}^{(1)}(T_i, \hat{\mathbf{B}})$, both of which include many individuals besides the ith one. Thus, Binder (1992) found an alternative expression for $\hat{U}_0(\hat{\mathbf{B}})$ which conforms to these conditions, making it possible to obtain a design consistent estimate $\hat{V}(\hat{U}_0(\hat{\mathbf{B}}))$ by application of a design-based variance estimation method to the alternate expression and then evaluating this variance estimate at $\mathbf{B} = \hat{\mathbf{B}}$. If the design-based variance estimation method chosen is the linearization method, then the first step consists of calculating the following residual for each of the sampled individuals:

$$\hat{u}_i(T_i, \hat{\mathbf{B}}) = \hat{u}_{i0}(T_i, \hat{\mathbf{B}}) - \sum_{j=1}^{N} w_j(s)\, \delta_j \frac{Y_i(T_j)\, e^{\mathbf{x}_i'(T_j)\hat{\mathbf{B}}}}{\hat{S}^{(0)}(T_j, \hat{\mathbf{B}})} \left\{ \mathbf{x}_i(T_j) - \frac{\hat{S}^{(1)}(T_j, \hat{\mathbf{B}})}{\hat{S}^{(0)}(T_j, \hat{\mathbf{B}})} \right\}. \tag{10}$$

Each individual in the sample belongs to a particular PSU within a particular stratum. Thus, instead of identifying an individual by a single subscript i we will use a triple subscript hci where $h = 1, 2, \ldots, H$ identifies the stratum, $c = 1, 2, \ldots, c_h$ identifies the PSU within the stratum and $i = 1, 2, \ldots, n_{hc}$ identifies the individual within the PSU. Then

$$\hat{V}(\hat{U}_0(\hat{\mathbf{B}})) = \sum_{h=1}^{H} \frac{1}{c_h(c_h - 1)} \sum_{c=1}^{c_h} (t_{hc} - \bar{t}_h)\, (t_{hc} - \bar{t}_h)',$$

where

$$t_{hc} = c_h \sum_{i=1}^{n_{hc}} w_{hci}\, \hat{u}_{hci} \quad \text{and} \quad \overline{t_h} = \sum_{c=1}^{c_h} t_{hc} \Big/ c_h.$$

3. Inference for multiple-spell hazard rate models

3.1 Model-based inference

If more than one spell is observed for an individual, it is realistic to assume that these spells are not independent. Thus, the partial likelihood function (2) is misspecified for multiple spells since it does not account for intra-individual correlation of the spells observed on the same individual. Following Lin and Wei (1989), it is sufficient to modify only the covariance matrix of the estimated model parameters since the correlated durations affect the variance while the model parameters can be estimated consistently without accounting for this correlation. Lin (1994) demonstrates how the covariance matrix of the estimated model parameters might be estimated when there is intra-individual correlation of spells, provided that spells from different individuals are independent.

3.2 Design-based inference

In a longitudinal survey with a multi-stage design, the multiple events can be correlated at different levels: the spells are clustered within an individual, and individuals are clustered within high-stage units. The positive intracluster correlation at any level adds extra variation to estimates calculated from such data, beyond what is expected under independence. The assumption of independence of observations when they are cluster-correlated leads to underestimating the true standard errors, which inflates the values of test statistics, and ultimately results in too-frequent rejection of null hypotheses. Thus, for multiple spells for individuals, where the data are from a longitudinal survey, accounting just for correlation within individuals is insufficient.

Design-based variance estimation for nested cluster-correlated data can be greatly simplified when it is reasonable to assume that individuals from different primary sampling units (PSU's) are uncorrelated. This is equivalent to assuming that the PSU's are sampled with replacement. This assumption also holds approximately when the first stage units are obtained by sampling without replacement, provided that the sampling rate at the first stage is very small. In such a case, an estimate of the between-PSU variability captures the variability among units in all subsequent stages, regardless of the dependence structure among observations within each PSU. For a recent summary of robust variance estimation for cluster-correlated data see Williams (2000). This implies that Binder's (1992) approach for robust variance estimation of the single-spell model in the case of a survey design having with-replacement sampling at the first stage can be directly applied to the multiple spell situation since it accounts for the impact of cluster-correlation at all levels within each PSU.

4. Three models for multiple spells

In order to allow the covariates to have different effects for spells of different orders, as well as to allow different time dependencies (baseline hazards), we are exploring three models for multiple spells. The models differ according to the definition of the risk set and the assumptions about the baseline hazard. Two of these models account for the order of the spells.

It should be noted, however, that in our work, spell order refers only to spells occurring in the observation period from which the data are collected and not to the entire history of an individual (unless these two time periods coincide). For example, by the first spell we mean a first spell in the observation period although it may be a spell of some higher absolute order over the person's lifetime. This limitation implies a careful interpretation of any impact that spell order may have on covariate effects or on time dependency.

Model 1: In the first model, the risk set is carefully defined to take spell order into account in the sense that an individual cannot be at risk of completing the second spell before he completes the first, *etc*. This model, known as the conditional risk set model, was proposed by Prentice, Williams and Peterson (1981) and was reviewed by Lin (1994). It was also discussed by Hamerle (1989) and Blossfeld and Hamerle (1989) in the context of modelling multi-episode processes. Generally, the conditional risk set at time t for the completion of a spell of order j consists of all individuals that are in their j^{th} spells. This model allows spell order to influence both the effect of covariates and the shape of the baseline hazard function.

The hazard function for the ith individual for the spell of jth order is

$$h_j(t \mid \mathbf{x}_{ij}(t)) = \lambda_{0j}(t)\, e^{\mathbf{x}'_{ij}(t)\boldsymbol{\beta}_j},$$

where, for each spell order, a different baseline hazard function and a different coefficient vector are allowed. For this model and for other models that will be considered in this Section, time t is measured from the beginning of the jth spell. Although spells within the same individual may not be independent, the following partial likelihood is still valid for estimation of the $\boldsymbol{\beta}_j$'s:

$$L(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K) = \prod_{j=1}^{K} \prod_{i=1}^{N_j} \left[\frac{e^{\mathbf{x}'_{ij}(T_{ij})\boldsymbol{\beta}_j}}{\sum_{r=1}^{N_j} Y_{rj}(T_{ij}) e^{\mathbf{x}'_{ij}(T_{ij})\boldsymbol{\beta}_j}} \right]^{\delta_{ij}}, \quad (11)$$

Here, $T_{1j}, ..., T_{N_j j}$ are N_j durations of possibly right-censored j^{th} order spells, $\delta_{ij} = 1$ if T_{ij} is an observed duration and $\delta_{ij} = 0$ otherwise, and K is the highest order of spells to be included in the Cox model. The denominator sum is taken over the jth spells that are at risk of being completed at time T_{ij}, *i.e.*, $Y_{rj}(t) = 1$ if $t \le T_{rj}$, and is equal to 0 otherwise. The corresponding covariate vector observed on $[0, T_{ij}]$ is $\mathbf{x}_{ij}(t)$. Partial likelihood (11) can be maximized separately for each j if there are no additional restrictions on the $\boldsymbol{\beta}_j$'s.

The corresponding score equations that define the finite population parameter $\mathbf{B} = (\mathbf{B}'_1, \mathbf{B}'_2, ..., \mathbf{B}'_K)'$ are:

$$U_0(\mathbf{B}) = \sum_{j=1}^{K} \sum_{i=1}^{N_j} u_{ij0}(T_{ij}, \mathbf{B}_j) = 0, \quad (12)$$

with

$$u_{ij0}(T_{ij}, \mathbf{B}_j) = \delta_{ij} \left\{ \mathbf{x}_{ij}(T_{ij}) - \frac{S^{(1)}(T_{ij}, \mathbf{B}_j)}{S^{(0)}(T_{ij}, \mathbf{B}_j)} \right\},$$

and with $S^{(0)}(t, \mathbf{B}_j)$ and $S^{(1)}(t, \mathbf{B}_j)$ having the form of (5) and (6) respectively, but with N_j replacing n and $\mathbf{B}_j$ replacing $\boldsymbol{\beta}$.

The design-based estimates of the parameters $\mathbf{B}_j$ are obtained by solving equations $\sum_{i=1}^{N_j} w_i(s) \hat{u}_{ij0}(T_{ij}, \hat{\mathbf{B}}_j) = 0$ separately for each j, where $\hat{u}_{ij0}$ has the form of u_{ij0} but with $S^{(0)}$ and $S^{(1)}$ replaced by $\hat{S}^{(0)}$ and $\hat{S}^{(1)}$ respectively. Note that the sampling weights correspond to individuals and not to spells. Similarly, estimation of the covariance matrix of each $\hat{\mathbf{B}}_j$ will be done separately using the design-based robust estimation approach described in Section 2.2. Technically, this is a set of analyses separated by spell order.

Model 2: The second model considered is the marginal model (Wei, Lin and Weissfeld 1989):

$$h_j(t \mid \mathbf{x}_{ij}(\mathrm{t})) = \lambda_{0j}(t)\, e^{\mathbf{x}'_{ij}(t)\boldsymbol{\beta}},$$

where, for each spell order, we allow a different baseline hazard function while the covariate effects are kept the same over different spell orders. The corresponding partial likelihood function as well as the risk set, under the assumption that spells within the same individual are independent, is the same as for Model 1, with $\boldsymbol{\beta}$ replacing the $\boldsymbol{\beta}_j$'s. The corresponding score equation that defines the finite population parameter is

$$U_0^*(\mathbf{B}) = \sum_{j=1}^{K} \sum_{i=1}^{N_j} u^*_{ij0}(T_{ij}, \mathbf{B}) = 0,$$

with

$$u^*_{ij0}(T_{ij}, \mathbf{B}_j) = \delta_{ij} \left\{ \mathbf{x}_{ij}(T_{ij}) - \frac{S^{(1)}(T_{ij}, \mathbf{B})}{S^{(0)}(T_{ij}, \mathbf{B})} \right\},$$

where $S^{(0)}(t, \mathbf{B})$ and $S^{(1)}(t, \mathbf{B})$ are defined by (5) and (6) respectively, but with N_j replacing n and $\mathbf{B}$ replacing $\boldsymbol{\beta}$.

The design-based estimate of the parameter $\mathbf{B}$ is obtained by solving the weighted score equations

$$\sum_{i=1}^{K} \sum_{j=1}^{N_i} w_i(s) \hat{u}^*_{ij0}(T_{ij}, \hat{\mathbf{B}}) = 0,$$

where $\hat{u}^*_{ij0}$ has the form of u^*_{ij0} but with $S^{(0)}(t, \mathbf{B})$ and $S^{(1)}(t, \mathbf{B})$ replaced by $\hat{S}^{(0)}(t, \hat{\mathbf{B}})$ and $\hat{S}^{(1)}(t, \hat{\mathbf{B}})$ respectively.

The estimation of the covariance matrix of $\hat{\mathbf{B}}$ will be done using the design-based robust estimation approach explained in Section 3.2.

Model 3: The last model considered is the following:

$$h_j(t \mid \mathbf{x}_{ij}) = \lambda_0(t)\, e^{\mathbf{x}'_{ij}(t)\boldsymbol{\beta}}.$$

In this model we assume that the baseline hazard functions and the effects of covariates are common for different orders of spells. The risk set at time T_{ij} is defined differently than for Models 1 and 2, and contains all spells with $t \le T_{ij}$, effectively assuming that all spells are from different individuals. Technically, this model is a single-spell model, so that estimation of coefficients and variances by a design-based robust method is straightforward.

5. Example of modelling multiple unemployment spells

5.1 The data

The data set that we use for illustration comes from the first six-year panel (1993-1998) of the Canadian Survey of Labour and Income Dynamics (SLID). In this panel, about 31,000 individuals from approximately 15,000 households were followed for six years through annual interviews. Some individuals dropped out of the sample over time for any number of reasons while a few others, after missing one or more interviews, resumed their participation. A complex weighting of the responding SLID individuals each year takes into account different types of attrition so that each respondent in a particular year is weighted against the

relevant reference population of 1993. This results in a separate longitudinal weight for each wave (*i.e.*, year) of data. For this analysis we used the longitudinal weights from the last year of the panel, *i.e.*, 1998, which meant that data just from those individuals who were respondents in the final wave of the panel were included in the analyses. A good summary of the sample design issues in SLID is given in Lavigne and Michaud (1998). A review of the issues related to studies of unemployment spells from SLID is given in Roberts and Kovačević (2001).

The state of interest is "being unemployed", defined in this case as the state between a permanent layoff from a full-time job and the commencement of another full-time job. A job is "full-time" if it requires at least 30 hours of work per week. The event of interest is "the exit from unemployment". Only spells beginning after January 1, 1993 were included since January 31, 1993 is the starting date for observations from the panel. Spells that were not completed by the end of the observation period (December 31, 1998) were considered censored. Sample counts of the number of individuals experiencing eligible spells and the number of spells according to their order are given in Table 1. In brief, there were17,880 spells from 8,401 longitudinal individuals. About half of the sampled individuals (4,260) who became unemployed during this period experienced two or more unemployment spells. There were 3,809 spells that remained uncompleted due to the termination of the panel.

From a long list of available covariates we chose only ten. The variable sex [SEX] of the longitudinal individual is the only variable that remains constant over different spells. Four variables have values recorded at the end of the year in which the spell commenced: education level [EDUCLEV] with 4 categories (low, low-medium, medium, high), marital status [MARST] with three categories (single, married/common law, other), family income per capita (in Canadian dollars) with 4 categories (<10K, 10K-20K, 20K-30K, 30K+), and age [AGE] (in years). Three variables have the values from the lay-off job preceding the spell: type of job ending [TYPJBEND] with two categories (fired and voluntary), occupation [OCCUPATION] with 6 categories (professional, administration, primary sector, manufacturing, construction, and others); and firm size [FIRMSIZE] with five categories (<20, 20-99, 100-499, 500-999, 1,000 + employees). Two binary variables represent the situation during the spell: having a part time job [PARTTJB], and attending school [ATSCH].

The data set was prepared in the "counting process" style where each individual with eligible spells is represented by a set of rows, and each row corresponds to a spell. Although a row contains time of entry to the spell t_1, and time of exit t_2 or time of censoring t_c, the duration time for analysis is always considered in the form $(0,\ t_2 - t_1)$ or $(0,\ t_c - t_1)$. The covariates under consideration are attached to each row. Also attached to each row are the 1998 longitudinal weight and the identifiers for the stratum and the PSU of the person whose spell is being described by that record.

Table 1 Counts of individuals in the six-year panel of SLID with unemployment spells beginning between January 1993 and December 1998, by the total number of spells and by order of spell (C-completed, U-uncompleted)

Individuals by number of spells		Spells by order									
		First		Second		Third		Fourth		5th +	
		C	U	C	U	C	U	C	U	C	U
1 spell	4,141	2,221	1,920	-	-	-	-	-	-	-	-
2 spells	1,915	1,915	-	1,154	761	-	-	-	-	-	-
3 spells	1,044	1,044	-	1,044	-	612	432	-	-	-	-
4 spells	629	629	-	629	-	629	-	348	281	-	-
5+ spells	672	672	-	672	-	672	-	672	-	1,158	415
Total	8,401	6,481	1,920	3,499	761	1,913	432	1,020	281	1,158	415

5.2 Analysis

For the purpose of this illustration we restricted the analysis to the first four spells, which means that all sampled individuals with eligible spells are included in the analysis but the spell records after the fourth spell are not considered due to their small number in the sample.

We estimated coefficients and their variances for the 3 models by the design-based methods described in Section 4 through the use of the "SURVIVAL" procedure in SUDAAN Version 8. For all three models, the survey design was specified to be stratified with with-replacement selection of PSU's (*i.e.*, DESIGN = WR). All three models were fit to the same number of spells (16,307). For each model, we then calculated the empirical cumulative baseline hazard functions using a product-limit approach (see Kalbfleisch and Prentice (2002), pages 114-116) as implemented in the SURVIVAL procedure in SUDAAN.

In the robust model-based approach for multiple spells described in Section 3.1, there is an adjustment in the variance estimates to account for the possible dependence among spells from the same individual, assuming independence of spells from different individuals; however, in this approach, no account is made for the unequal probabilities of selection of the sampled individuals - in either the coefficient estimates or the variance estimates. In order to do this, for models 1 and 2 we also used the SURVIVAL procedure in SUDAAN Version 8, to estimate the variances of the weighted coefficient estimates where we assumed independence of spells between individuals but allowed for possible correlation of spells from the same individual. We did this by specifying the sampling design to be unstratified and having with-replacement selection of clusters, and we specified that each individual formed his own cluster. The dependence assumptions are the same as those used by Lin (1994) but we accounted for the use of weights in the estimation of the coefficients and the variances. We will call these variance estimates "modified robust model-based variance estimates of weighted coefficient estimates".

5.3 Some descriptive statistics

The estimated mean duration of a completed spell is 33.3 weeks while the estimated mean duration of the observed portion of a censored (uncompleted) spell is 48.5 weeks.

Visual examination of estimated Kaplan-Meier survival functions (not shown) for spells of each order indicated that, as order increased, the value of the survivor function at any fixed time t decreased, indicating that first spells are the longest among completed spells, and that the higher the order of a multiple spell the shorter is its duration. This is likely to be a consequence of the limited life of the panel, in the sense that an individual with more spells in the given six-year time frame is likely to have shorter spells.

5.4 Model fits using a design-based approach

As noted earlier, our example is just an illustration of the design-based approach to fitting proportional hazards models to multiple-event data from a survey with a complex design. Thus, little time is spent in this article on discussing how to assess the adequacy of these models, such as the adequacy of the proportionality assumptions in all of the models or whether one type of model fits as well as another.

Estimated coefficients from fitting the three models to the SLID data are given in Table 2. Coefficients found significant at the 5% level, through the use of individual t tests, are shown in bold.

Model 1 is conditional on the spell order and involved fitting four models separately to the data from the four different spell orders. As seen in Table 2, SEX, AGE, and at least one category of the Family Income variable were significant for spells of all orders, although magnitudes of the estimated coefficients differed with the spell order. The estimated coefficients for AGE were negative but decreased in magnitude as the spell order increased, while there was no discernable pattern in the estimated coefficients for the other 2 variables. The variables EDUCLEV, PARTJB and ATSCH had significant coefficients for spells of order 1, 2, and 3, but not for spells of order 4. This can be at least partly attributed to the small sample size for the fourth spells. For each of the other three variables in the model (MARST, OCCUPATION, and FIRMSIZE), there was just one spell order for which a coefficient was significant.

For Model 2, the model coefficients are restricted to be the same for all spell orders. As seen in Table 2, numerically many - but not all - of the estimated coefficient values were situated between the estimates for the first and the second spells obtained for Model 1 which could be due to the fact that a high proportion of the sample corresponded to events of these orders. All but the OCCUPATION variable had a significant coefficient. Standard errors of coefficients were smaller for Model 2 than for Model 1.

Model 3 is a single-spell model with a single set of model coefficients and a single baseline hazard function. The estimated model coefficients are similar to the estimates obtained by Model 2.

Table 2 Estimated β coefficients for three models

	Model 1				Model 2	Model 3
	Order 1	**Order 2**	**Order 3**	**Order 4**		
SEX (F)						
M	**0.4417**	**0.3781**	**0.3299**	**0.4435**	**0.4049**	**0.4090**
EDUCLEV (H)						
L	**-0.4561**	**-0.5234**	**-0.3748**	-0.1065	**-0.4128**	**-0.4331**
LM	**-0.2330**	**-0.2700**	**-0.3310**	-0.1653	**-0.2436**	**-0.2474**
M	-0.0744	-0.1060	-0.1156	0.0668	-0.0684	-0.0671
MARST (M)						
Single	**-0.1142**	-0.1290	-0.0622	-0.1375	**-0.1357**	**-0.1330**
Other	0.0985	-0.0894	0.1124	-0.1072	0.0328	0.0401
TYPJBEND (Fired)						
Voluntary	0.0704	**0.2752**	**0.4207**	**0.3413**	**0.1579**	**0.1284**
OCCUPATION(Othrs)						
Professionals	0.1592	-0.1364	-0.1388	0.0903	0.0490	0.0485
Admin	-0.0265	**-0.2930**	-0.1769	0.0579	-0.0971	-0.0938
PrimSector	-0.0211	-0.2175	-0.1187	0.2032	-0.0410	-0.0201
Manufacture	-0.0003	-0.0994	-0.1295	0.2862	-0.0093	-0.0088
Construction	0.1290	-0.1862	-0.0879	0.2339	0.0490	0.0813
FIRMSIZE (1000+)						
<20	-0.0027	-0.0097	0.1005	**0.4403**	0.0441	0.0408
20-99	0.0358	0.0881	0.0815	0.3999	**0.0928**	**0.0951**
100-499	0.0436	-0.0905	0.0328	0.0257	0.0214	0.0278
500-999	-0.0006	0.0153	-0.0623	-0.0067	-0.0005	0.0020
PARTTJB (No)						
Yes	**-0.2903**	**-0.5414**	**-0.5109**	-0.1407	**-0.3693**	**-0.3743**
ATSCH (No)						
Yes	**-1.0832**	**-1.1516**	**-1.2956**	-1.3541	**-1.1205**	**-1.1266**
Family Income Per Capita (10K-)						
10K-20K	**0.1294**	**0.1802**	0.0692	**0.1117**	**0.1345**	**0.1330**
20K-30K	**0.1644**	**0.3611**	0.1572	**0.4900**	**0.2241**	**0.2141**
30K+	**0.1712**	**0.3916**	**0.3005**	**0.4241**	**0.2280**	**0.2115**
AGE	**-0.0491**	**-0.0311**	**-0.0269**	**-0.0207**	**-0.0424**	**-0.0435**
Spells in risk set	8,386	4,255	2,345	1,300	16,286	16,286
Censored	1,913	759	432	281	3,385	3,385
Completed	6,473	3,496	1,913	1,019	12,901	12,901

The values significant at a 5% level are bold.

The estimated cumulative baseline hazard functions for Models 1 to 3 are given in Figures 1 to 3 respectively. In all cases, for durations up to approximately 50 weeks, the functions have a concave shape, implying that there is a positive time dependence of the exit rate (*i.e.* the longer the spell, the higher the probability of exit). For durations longer than 50 weeks, the shapes become convex, suggesting negative time dependence for the longer spells. In Figure 1, positions of the estimated cumulative baseline hazard functions vary according to spell order, with the curve for spells of order 1 being the highest, and the curve for spells of order 4 being the lowest. In Figure 2, for Model 2, the positions of the different curves do not follow spell order. This observed difference between Figures 1 and 2 could serve as one visual diagnostic that further study is required in order to assess whether Model 1 or Model 2 is a better descriptor of the data, since estimated coefficients have an impact on the estimated baseline hazards.

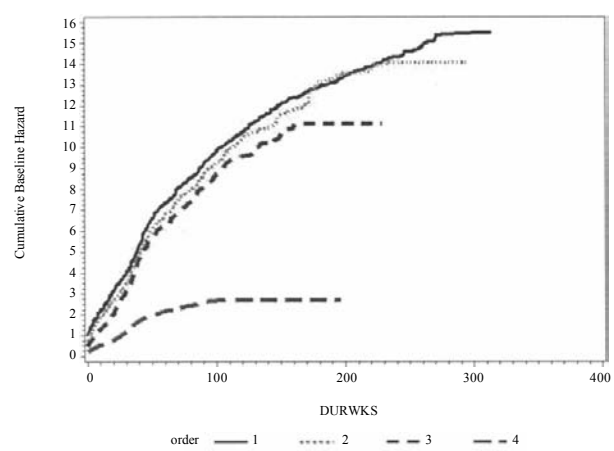

Figure 1 Cumulative Baseline Hazard – Model 1

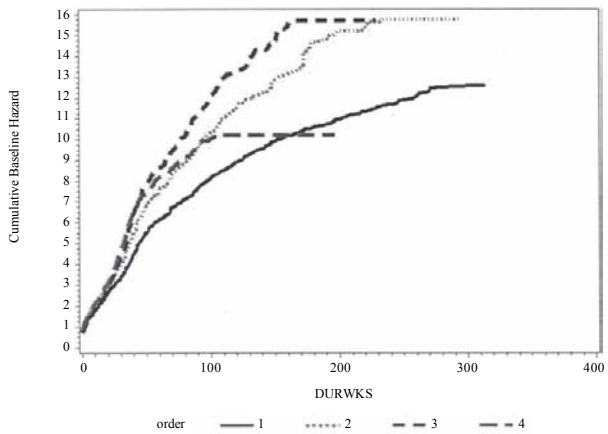

Figure 2 Cumulative Baseline Hazard – Model 2

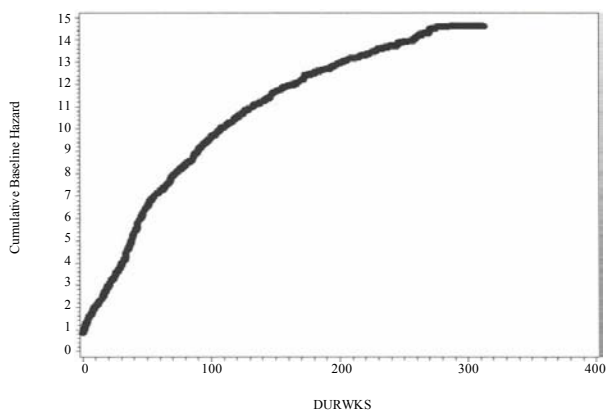

Figure 3 Cumulative Baseline Hazard – Model 3

5.5 Comparison to modified robust model-based variance estimates

As described in Section 5.2, the modified robust model-based variance estimates account for possible correlation among spells from the same individual, where independence among individuals is assumed. When, for Models 1 and 2, the standard error estimates obtained by this approach were compared to the design-based standard error estimates, only very minor differences were observed. This would seem to indicate that the design-based estimates are picking up any correlation among spells from the same individual and also that there does not appear to be additional dependence above the level of the individual for our particular example.

6. Concluding remarks

We explored the problem of analysis of multiple spells by considering two general approaches for dealing with the lack of independence among the exit times: a robust model-based approach and a design-based approach. The first approach estimates the model parameters assuming independence of the spells, and then corrects the naïve covariance matrix to account for within-individual dependencies postulated by the researcher. This approach does not account for the possible clustering between individuals (or, in fact, for any clustering that might occur at a level above the individual) nor for the unequal probabilities of selection of individuals (although, in our example, we showed how the method could be extended to include the survey weights). The second approach defines the model coefficients as finite population parameters. These parameters are then estimated accounting for possible unequal selection probabilities of individuals. A design-based variance estimation method that accounts for possible correlations between individuals in the same PSU automatically accounts for the unspecified dependencies of spells at levels below the PSU, such as dependencies within individuals. For large sample sizes this design-based inference extends directly to the super-population from which, hypothetically, the finite population was generated. The deficiency of the first approach is that it totally ignores the potential for clustering between individuals. A possible disadvantage of the second approach, as we applied it, is that it relies on the assumption of with-replacement sampling of PSU's of individuals. The two approaches coincide in the case of simple random sampling of individuals where, in the robust model-based approach, dependence among spells from the same individual is explicitly postulated and accounted for in the variance estimation formula and where, in the design-based approach, spells from the same individual are treated as a cluster in the design-based variance estimation.

We applied the design-based approach to three proportional-hazards-type models. One model allowed for differential unspecified baseline hazards and different coefficients for each spell order. The second model still allowed for differential unspecified baseline hazards for different spell orders but required the coefficients to be the same over orders. The third model was a simple single-spell model. We found that how information on the spell order was used affected the results of our model-fitting. A visual compareison of the coefficient estimates and the estimates of the cumulative baseline hazards for Models 1 and 2 indicated different results. A formal test for whether the coefficients actually differ by spell order (as allowed in Model 1), given baseline hazards that can differ by spell order, would be useful, as suggested by one of the referees. It is actually straightforward to produce such a test, and can be done as follows. Let $\boldsymbol{\gamma} = (\mathbf{B}_1', \mathbf{B}_2', ..., \mathbf{B}_K')'$ be the vector of all K coefficient vectors of Model 1, where each has length p, and let $\mathbf{z}_{ij}(t) = (0', 0', ..., \mathbf{x}_{ij}(t)', 0', ..., 0')'$ be the vector of length pK for the jth spell of the ith individual where the jth component of this vector contains the vector of covariates $\mathbf{x}_{ij}(t)$. Then, Model 1 can be expressed as

$$h_j(t \mid \mathbf{z}_{ij}(\mathrm{t})) = \lambda_{0j}(t)\, e^{\mathbf{z}_{ij}'(t)\boldsymbol{\gamma}},$$

which has the general form of baseline hazards varying with spell order but a fixed coefficient vector. A test for constancy of the coefficients pertaining to each spell order, *i.e.*, $H_0 : \mathbf{B}_1 = \mathbf{B}_2 = \ldots \mathbf{B}_K$ is equivalent to testing $H_0 : \mathbf{C}\boldsymbol{\gamma} = 0$ where $\mathbf{C}$ is the $(K-1)p \times Kp$ matrix $\mathbf{C} = I_p \otimes [1_{K-1} \ -I_{K-1}]$. Given an estimate $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$ and an estimate $\hat{V}(\hat{\boldsymbol{\gamma}})$ of the covariance matrix of $\hat{\boldsymbol{\gamma}}$, obtained as described in Section 4 for Model 2, a Wald statistic may be calculated in order to test the hypothesis. If the hypothesis is not rejected, it may be concluded that a model with constant coefficients over spell order (but baseline hazard varying with spell order) appears to fit the data as well as a model where both baseline hazard and coefficients vary with spell order. Other measures for model adequacy should also be straightforward to develop under the design-based framework.

We also visually compared, for our example, coefficient standard error estimates obtained under the design-based approach (accounting for clustering at the PSU level and lower) and obtained under a modification of the robust model-based approach (accounting for clustering at the individual level and lower) for Models 1 and 2. We found only minor differences, which indicated no clustering effects above the individual level for these particular data. We also calculated standard error estimates assuming independence even between spells from the same person and again found only minor differences with those obtained from the design-based approach. It thus seems that, for this particular example, there is little inter-spell dependence. However, in general, we feel that a design-based approach guards against missing any unpostulated dependencies at the PSU level and lower in the variance estimates.

Acknowledgements

We are grateful to Normand Laniel and Xuelin Zhang for their useful comments to an earlier version of this manuscript. We also thank the associate editor and the referees for comments and suggestions improving greatly the readability of the manuscript.

References

Binder, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-291.

Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika,* 79, 139-147.

Blossfeld, H.-P., and Hamerle, A. (1989). Using Cox models to study multiepisode processes. *Sociological Methods and Research*, 17, 4, 432-448.

Clayton, D., and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society*, Series A, 1985, 148, 82-117.

Cochran, W.G. (1977). *Sampling Techniques*. Third edition. New York: John Wiley & Sons, Inc.

Cox, D.R. (1975). Partial likelihood. *Biometrika*, 62, 269-276.

Hamerle, A. (1989). Multiple-spell regression models for duration data. *Applied Statistics*, 38, 1, 127-138.

Heckman, J., and Singer, B. (1982). Population heterogeneity in demographic models. In *Multidimensional Mathematical Demography*, (Eds., K. Land and A. Rogers), New York: Academic Press, 567-599.

Hougaard, P. (1999). Fundamentals of survival data. *Biometrics*, 55, 1, 13-22.

Kalbfleisch, J.D., and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd Edition, New York: John Wiley & Sons, Inc.

Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48, 795-806.

Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica*, 47, 939-956.

Lavigne, M., and Michaud, S. (1998). General Aspects of the Survey of Labour and Income Dynamics. Working Paper, Statistics Canada, 75F0002M No. 98-05.

Lin, D.Y. (1994). Cox regression analysis of multivariate failure time data: a marginal approach. *Statistics in Medicine*, 13, 2233-2247.

Lin, D.Y. (2000). On Fitting Cox's proportional hazards models to survey data. *Biometrika*, 87, 37-47.

Lin, D.Y., and Wei, L.J. (1989). The robust Inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84, 1074-1078.

Prentice, R.L., Williams, B.J. and Peterson, A.V. (1981). On the regression analysis of multivariate failure data. *Biometrika*, 68, 373-379.

Roberts, G., and Kovačević, M. (2001). New research problems in analysis of duration data arising from complexities of longitudinal surveys. *Proceedings of the Survey Methods Section of the Statistical Society of Canada,* 111-116.

Wei, L.J., Lin, D.Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84, 1065-1073.

Williams, R.L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics*, 56, 645-646.