MODELLING MULTIPLE UNEMPLOYMENT SPELLS FROM LONGITUDINAL SURVEY DATA

Milorad S. Kovacevic (kovamil@statcan.ca) and Georgia Roberts (robertg@statcan.ca)
Statistics Canada, Ottawa, Canada, K1A 0T6

KEY WORDS: Cox Regression, Design-Based Inference, Model-Based Inference, Spell Order, SLID

1. INTRODUCTION

The modelling problem addressed in this paper has been called correlated failure-time modelling, multivariate survival modelling, multiple spells modelling, or a recurrent events problem, and is studied in biomedical (e.g., Lin, 1994, Hougaard, 2000), social (Blossfeld and Hamerle, 1989, Hamerle, 1989) and economic literature (Lancaster, 1979, Heckman and Singer,1982). Generally this type of modelling is required for data that arise in time-to-event studies when two or more events happen to the same subject. In such a case, the failure times are correlated within subject, and thus the assumption of independence of failure times conditional on given measured covariates, required by standard survival models, is violated. The research interest is usually to assess the effect of various covariates considered as potential risk factors.

In studies of duration of spells (poverty, jobless-ness, etc.), the 'failure' is equivalent to 'exit' out of the state of interest. The dependence among the observed spells from the same individual comes 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 seems to be 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.). The paper does not explore this method any further.

Another approach is to treat the dependence among multiple spells as a nuisance, and to model the marginal distributions of the individual spells without explicit modelling of the dependencies among the spells, with a possible utilization of the order of the spells in the model specification. Following Lin and Wei (1989) it is possible to modify only the 'naive' covariance matrix of the estimated model parameters 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 parameters per se. An additional property of many multiple spells, often ignored, is that they are ordered 'events': the second spell cannot occur before the first. This approach of working independence and corrected variance can easily be made to account for the order of the spells.

In socio-economic studies of duration of spells the data usually come from 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. There is a need to account for the sample design when estimating the model parameters and the variances of these estimates. One of the early references for the use of complex sample data for estimation of proportional hazards models is Chambless and Boyle (1985). They estimated the discrete proportional hazards model introduced by Prentice and Gloecker (1978) by solving the likelihood score equations weighted by the sampling weights. This method is known in survey sampling literature as the pseudo-likelihood method (Skinner, 1989). In order to estimate the standard errors, they first verified the asymptotic normality of the weighted likelihood estimates, and then applied Binder's (1983) method for the design-based estimation of the variances of asymptotically normal estimators. Binder (1992) examines the fitting of a continuous proportional hazards model to survey data by first defining the finite population parameter of interest as the solution to the partial likelihood score equations based on the entire finite population and then estimating the finite population parameters and the variance of these estimates by applying Binder's method (1983). In order to estimate the variance he used an alternative expression of the partial likelihood score equation derived by Lin and Wei (1989) for independent sampling. Recently, Lin (2000) provided formal justification of Binder's (1992) method and extended it to the superpopulation framework where inference accounts for both sources of randomness, one generated by the assumed model and the other coming from the sample design.

Our approach is to model the marginal distributions of the multiple spells using single spell models, treating the dependence among the spells as a nuisance. The finite population parameters are defined as a solution of the resulting partial likelihood score equations and these parameters are estimated using design-based estimation. The covariance is estimated using an appropriate design consistent linearization method assuming that the primary sampling units are sampled with replacement within strata. This assumption is viable given the small sampling rates usually used 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 the methods for robust estimation of the variances when the model is misspecified. Section 3 contains further discussion on robust variance estimation for multiple spells. Then, in Section 4, three models are introduced and applied to the multiple spells. A full description of how to fit these models using the single-spell robust estimation methods is also given. In Section 5 we fit these models to the data from the Canadian Survey on Labour and Income Dynamics (SLID) and discuss the numerical results. Finally, Section 6 contains some overall remarks.

2. DISTRIBUTION OF SINGLE SPELLS AND THE STANDARD HAZARD RATE MODEL

The duration of a spell (or simply, a spell) experienced by an individual is a random variable denoted by *T*. The main characteristics of *T* such as the cumulative distribution $F(t)$, probability density $f(t)$, expectation μ, etc., may be defined in the usual way. For the spell T, however, we are more interested in quantities such as: (i) the survivor function of T, $S(t) = Prob\{T>t\} = 1 - F(t)$, defined as the probability that the spell is not completed at *t*, and (ii) the hazard function of *T* at *t*, *h(t)*,defined as the probability that the spell is completed at *t* given that it has not been completed before *t*,

$$h(t) = \lim_{dt \to 0} \frac{Prob\ \{t \leq T<t+dt \mid T \geq t\}}{dt}$$

The value of the hazard function at *t* is called the exit rate to emphasis that the completion of the spell is equivalent to exit out of the state of interest. The hazard function provides a full characterization of the distribution of *T*, just like $F(t)$, $f(t)$, or $S(t)$. In other words, once the hazard function is completely specified the distribution of the spell is also completely specified. 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, the main concern is usually to study the impact of some key covariates on the distribution of *T*. We assume that the variation in distribution of spells can be characterized by a vector of observed explanatory variables $\boldsymbol{x}$ which can be time-invariant or time-varying variables. Under the proportional hazards model, the hazard function for the spell *T* associated with a vector of possibly time-varying covariates $\boldsymbol{x} = (x_1,...,x_p)'$ is

$$h(t|\boldsymbol{x}) = \lambda_0(t)\, e^{\,x'(t)\,\beta} \qquad (1)$$

The function $\lambda_0(t)$ is an unspecified baseline hazard function and gives the shape of the hazard function. If an individual has all x variables set at 0, the value (level) of the hazard function is equal to the baseline hazard. Similarly, if two individuals have identical values of the observed characteristics, they have identical hazard functions (1). The baseline hazard describes the duration dependance, namely whether the hazard rate depends on time already spent in the spell. For example, negative dependance describes the situation where the longer the spell the smaller the probability of exit.

Vector β contains the unknown regression parameters showing the dependance of the hazard on the $\boldsymbol{x}$ variables and may be estimated by maximizing the partial likelihood function (Cox, 1975):

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{e^{\,x_i'(T_i)\,\beta}}{\sum_{j=1}^{n} Y_j(T_i)\, e^{\,x_j'(T_i)\,\beta}} \right]^{\delta_i} .$$

Here T_1,...T_n are n durations possibly right-censored; $\delta_i = 1$ if T_i is an observed duration and $\delta_i = 0$ otherwise; and $\boldsymbol{x}_i(t)$ 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., $Y_j(t) = 1$ if $t \leq T_j$, and is equal to 0, otherwise]. The estimate of the model parameter β is obtained by solving the partial likelihood score equation

$$U_0(\beta) = \sum_{i=1}^{n} u_{i0}(T_i, \beta) = 0 \qquad (2)$$

where $u_{i0}(T_i, \beta) = \delta_i \left\{ x_i(T_i) - \dfrac{S^{(1)}(T_i, \beta)}{S^{(0)}(T_i, \beta)} \right\}$ is the score residual, with $S^{(0)}(t, \beta) = \dfrac{1}{n} \sum_{i=1}^{n} Y_i(t)\, e^{\,x_i'(t)\beta}$ and $S^{(1)}(t, \beta) = \dfrac{1}{n} \sum_{i=1}^{n} Y_i(t)\, x_i(t)\, e^{\,x_i'(t)\beta}$. If the model (1) is true, the model-based variance matrix of the score function $U_0(\beta)$ is

$$J(\beta) = \partial U_0(\beta) / \partial\beta$$

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

where $S^{(2)}(t, \beta) = \dfrac{1}{n} \sum_{i=1}^{n} Y_i(t)\, x_i(t) x_i'(t) e^{\,x_i'(t)\beta}$. The approximate variance of $\hat{\beta}$, obtained by linearizarion, is $J^{-1}(\hat{\beta})$.

If the real dependance structure is misspecified by the model, Lin and Wei (1989) provide the robust variance estimator for $\hat{\beta}$ as

$$J^{-1}(\hat{\beta})\,G(\hat{\beta})\,J^{-1}(\hat{\beta}), \qquad (3)$$

where $G(\beta) = \sum_{i=1}^{n} g_i(\beta)\, g_i^{/}(\beta)$ and

$$g_i(\beta) = u_{i0}(T_i, \beta) - \sum_{j=1}^{n} \delta_j \frac{Y_i(T_j)\, e^{x_i^{/}(T_j)}}{S^{(0)}(T_j, \beta)} \left\{ x_i(T_j) - \frac{S^{(1)}(T_j, \beta)}{S^{(0)}(T_j, \beta)} \right\}. \qquad (4)$$

For estimation of the parameters of a proportional hazards model from clustered survey data in the case of a single spell per individual, Binder (1992) used the estimating equations method. In particular, he first defined the finite population parameter of interest as a solution of the partial likelihood score equation (2) calculated from the spells of the entire population

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

where $u_{i0}(T_i, B)$ is the score residual defined in the same way as $u_{i0}(T_i, \beta)$, as well as $S^{(0)}(t, B)$ and $S^{(1)}(t, B)$. An estimate $\hat{B}$ of the parameter B is obtained as a solution to the partial pseudo-score estimating equation

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

where $w_i(s) = w_i$ if $i \in s$, and 0 otherwise, and $w_i(s)$ denotes a scaled sampling weight ($\sum w_i(s) = 1$). Function $\hat{u}_{i0}(T_i, \hat{B})$ takes the form

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

where $\hat{S}^{(0)}(T_i, \hat{B})$ and $\hat{S}^{(1)}(T_i, \hat{B})$ are the estimates of the respective means defined previously.

Generally, the design-based variance of an estimate $\hat{B}$ which satisfies the estimating equation $\hat{U}(\hat{B}) = \sum w_i u_i(\hat{B}) = 0$ can be obtained using linearization as

$$\hat{J}^{-1}\, \hat{V}(\hat{U}(B))\, \hat{J}^{-1} \qquad (5)$$

where $\hat{J} = \partial \hat{U}(B) / \partial B$ is evaluated at $B = \hat{B}$, and $\hat{V}(\hat{U}(\hat{B}))$ is the variance of the total estimated by some standard variance estimation method, see for example Cochran (1976).

Binder (1983) gave the conditions that the $u_i(B)$ functions should satisfy in order to provide consistent estimates of the variances of the implicitly defined parameters. In the case above, $\hat{U}_0(\hat{B})$ does not satisfy these conditions since the u_{i0}'s are the functions of the sums that depend on T_i. Rewriting the score residuals in a different way, Binder (1992) provided a weighted form of residuals (4) and the $u_i(\hat{B})$ functions such that the resulting $\hat{U}(\hat{B})$ conforms to the conditions:

$$\hat{u}_i(\hat{B}) = \delta_i \left\{ x_i(T_i) - \frac{\hat{S}^{(1)}(T_i, \hat{B})}{\hat{S}^{(0)}(T_i, \hat{B})} \right\} - \sum_{j \in s} w_j \delta_j \frac{Y_i(T_j)\, x_i(T_j)\, e^{x_i^{/}(T_j)\hat{B}}}{\hat{S}^{(0)}(T_j, \hat{B})} \left\{ x_i(T_j) - \frac{\hat{S}^{(1)}(T_j, \hat{B})}{\hat{S}^{(0)}(T_j, \hat{B})} \right\}. \qquad (6)$$

Using these values it is possible to obtain a design consistent estimate $\hat{V}(\hat{U}(\hat{B}))$ by application of some of the design-based methods.

The design-based approach provides estimation of a well-defined quantity even when the model is misspecified. Note that for a simple random sample of individuals the variance estimator (5) reduces to (3).

3. ROBUST INFERENCE FOR THE MULTIPLE SPELL HAZARDS MODELS

If more than one spell is observed for an individual, it is realistic to assume that these spells are not independent. Thus the likelihood function based on model (1) 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 necessary and 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. This implies that the model parameters can be estimated by treating events as independent, and then the variance estimates can be modified to account for the dependencies. The modified variance estimator (3) is robust to this type of model misspecification.

Socio-economic data are usually collected using a multi-stage sample design with compact geographic areas called primary sampling units (PSU) (e.g., census enumeration areas, neighbourhoods, city blocks or villages, etc.) being sampled at the first stage, and individuals being subsampled at the final stage. In a longitudinal study the multiple events are observed on the same person. These data are cluster-correlated at two nested levels: the spells are clustered within an individual, and individuals are clustered within the PSUs. The positive intra-cluster correlation at any level adds an extra variation to estimates calculated from such data, beyond what is expected under independence. Assuming independence of observations when they are cluster-correlated leads to underestimation of the true standard errors, inflates the values of test statistics, and ultimately results in too frequent rejection of null hypotheses.

The design-based variance estimation for the nested-cluster-correlated data can be greatly simplified by assuming that individuals from different PSU's are independent. This is equivalent to assuming that the PSUs are sampled with replacement. This assumption holds in large samples when the sampling rate at the first stage is very small so that the probability of having the same PSU sampled twice is negligible. In such a case, the estimate of the between-PSU variance captures the variability among units in all subsequent stages, allowing for arbitrary dependence structure among observations within a cluster. For a recent summary of robust variance estimation for cluster-correlated data see Williams (2000).

This implies that Binder's (1992) approach for the robust variance estimation of the single spell models can be directly applied to the multiple spells situation since it accounts for the clustering at the PSU level, while individuals are within PSU's. Briefly, the estimate $\hat{B}$ is obtained by solving the corresponding partial pseudo score equation $\hat{U}_0(\hat{B}) = 0$ assuming the independence of spells. The estimated design-based covariance matrix of $\hat{B}$ is obtained as $\hat{J}^{-1}(\hat{B})\,\hat{V}\{\hat{U}(\hat{B})\}\,\hat{J}^{-1}(\hat{B})$ where $\hat{U}(\hat{B})$ is the design-based estimate of the total of the $\hat{u}_i(\hat{B})$ functions defined by (6) with the quantities $S^{(0)}(t,B)$ and $S^{(1)}(t,B)$ appropriately defined and estimated.

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 dependancies (baseline hazards) we are fitting three models to 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 the spell order refers only to 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 consider a first spell in the observation period although it may be a spell of any higher absolute order over the person's lifetime. This limitation implies a careful interpretation of any effect 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 the order of the spells into account in the sense that an individual cannot be at risk of completing the second spell before he completes the first. 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 i th individual for the spell of jth order is

$$h_j(t \mid x_{ij}) = \lambda_{0j}(t)\, e^{x_{ij}^{/}(t)\beta_j}$$

where for each spell order, a different baseline hazard function and a different coefficient vector are allowed. An appropriate partial likelihood for this model, pretending that the spells within the same individual are independent, is

$$L(\beta_1,...,\beta_K) = \prod_{j=1}^{K} \prod_{i=1}^{N_j} \left[\frac{e^{x_{ij}^{/}(T_{ij})\beta_j}}{\sum_{r=1}^{N_j} Y_{rj}(T_{ij})\, e^{x_{rj}^{/}(T_{ij})\beta_j}} \right]^{\delta_{ij}} \qquad (7)$$

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. The denominator sum is taken over the j-th 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. $\boldsymbol{x_{ij}(t)}$ is the corresponding covariate vector observed on [0, T_{ij}]. Partial likelihood (7) can be maximized separately for each j if there are no additional restrictions on β_j.

The corresponding score equations that define the finite population parameter are

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

with $u_{ij0}(T_{ij}, B_j) = \delta_{ij} \left\{ x_{ij}(T_{ij}) - \frac{S^{(1)}(T_{ij}, B_j)}{S^{(0)}(T_{ij}, B_j)} \right\}$ and

$$S^{(0)}(t, B_j) = \frac{1}{N_j} \sum_{i=1}^{N_j} Y_{ij}(t)\, e^{x_{ij}^{/}(t) B_j} \text{ and}$$

$$S^{(1)}(t, B_j) = \frac{1}{N_j} \sum_{i=1}^{N_j} Y_{ij}(t)\, x_{ij}(t)\, e^{x_{ij}^{/}(t) B_j}.$$

The design-based estimates of the parameters B_j are obtained by solving equations

$$\sum_{i=1}^{N_j} w_i(s)\, u_{ij0}(T_{ij}, \hat{B}_j) = 0 \text{ separately for each } j.$$

Note that the sampling weights correspond to the individuals and not to the spells. Similarly, estimation of the variances will be done separately for each spell order using the same robust estimator (5). 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)

$$h_j(t\,|\,\boldsymbol{x}_{ij}) = \lambda_{0j}(t)\, e^{x_{ij}^{\prime}(t)\beta}$$

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, pretending that the spells within the same individual are independent, is the same as for Model 1. The corresponding score equation that defines the finite population parameter is

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

with $\quad u_{ij0}(T_{ij},B) = \delta_{ij}\left\{x_{ij}(T_{ij}) - \frac{S^{(1)}(T_{ij},B)}{S^{(0)}(T_{ij},B)}\right\}\quad$ where

$S^{(0)}(t,B) = \frac{1}{N_j}\sum_{i=1}^{N_j} Y_{ij}(t)\, e^{x_{ij}^{\prime}(t)B}$ and

$S^{(1)}(t,B) = \frac{1}{N_j}\sum_{i=1}^{N_j} Y_{ij}(t)\, x_{ij}(t)\, e^{x_{ij}^{\prime}(t)B}$.

The design-based estimate of the parameter B is obtained by solving the weighted score equations

$$\sum_{j=1}^{K} \sum_{i=1}^{N_j} w_i(s)\, u_{ij0}(T_{ij},\hat{B}) = 0 .$$

The estimation of the variance will be done using the robust estimator (5). Technically, this is an analysis stratified by spell order.

Model 3. The last model considered is also a model of the marginal hazard function

$$h_j(t\,|\,\boldsymbol{x}_{ij}) = \lambda_{0}(t)\, e^{x_{ij}^{\prime}(t)\beta}$$

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 is defined differently than for Models 1 and 2, and contains all spells with $t \leq T_{ij}$, effectively assuming that spells come from different individuals. Technically, this model is a single-spell model.

5. EXAMPLE WITH SLID DATA AND DISCUSSION

The data set that we use for illustration, comes from a six-year panel (1993-1998) of the Canadian Survey of Income and Labour Dynamics (SLID). About 31,000 longitudinal individuals were followed for six years with some being lost over time for any number of reasons. Some individuals lost to one or more interviews were found and resumed their participation. A complex weighting of the responding SLID individuals each year takes into account different types of attrition so that each respondent 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. A good summary of the sample design issues in SLID is given in Lavigne and Michaud (1998). A good review of the issues related to studies of unemployment spells from SLID is given in Roberts and Kovacevic (2001).

The state of interest is “unemployment” defined, for our example, as the state after permanent layoff from a full-time job until another full-time job begins. 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. Spells that were not completed by the end of the observation period (December 31, 1998) were considered as censored. Sample counts of the number of individuals experiencing spells and the number of spells of each order are given in Table 1. The data file used for this illustration contains 17,880 layoff records from 8401 longitudinal individuals. Evidently, about half of the sampled individuals (4260) who had spells experienced two or more unemployment spells. There are 3394 spells that remained uncompleted due to the end of the panel.

The data set for the illustration is prepared in the “counting process” style where each respondent 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 of interest are attached to each row. Some of them are describing the respondent (e.g., sex, age, and income), some of them are related to the job that ended when the spell commenced (e.g. firm size), and some are characteristics of the spell (e.g. receipt of employment insurance during the spell, or going back to school during the spell). Since some of these variables were recorded only once a year, for this illustration we used their values from the year in which the spell ended or was censored. Some covariates stayed constant over the life of the panel (e.g., sex), and others changed from spell to spell. Also attached to each row, for the purposes of point estimation and variance estimation, are the person longitudinal weight, and identifiers for the stratum and psu of the person whose spell is being described by that record.

We now apply the models and techniques described in the previous sections to the SLID data set. We used SAS and SUDAAN for carrying out our computations. For the purpose of this illustration we restricted the analysis to the first four spells, which meant that all sampled individuals with spells would be included in the analysis but that records for spells after the fourth were not considered. The estimated average duration of completed spells is 33.3 weeks while the average duration for censored (uncompleted) spells is 48.5 weeks.

Visual examination of estimated survival functions (not shown) indicated that, as order increased, the value of the SDF at any fixed time t decreased, indicating that single spells are the longest among completed spells, and that the higher the order of a multiple spell the shorter is its duration. This is a direct consequence of the limited life of the panel, so that an individual with more spells is likely to have shorter spells.

From a long list of available covariates we chose only ten. The variable SEX of the longitudinal individual is the only variable that remains constant over different spells. Other variables have values recorded at the end of the year in which the spell commenced (education level [EDUCLEV], marital status [MARST], family income per capita, age) or they have the values from the lay-off job preceding the spell (type of job ending[TYPJBEND], occupation, firm size) or they represent the situation during the spell (having a part time job[PARTTJB], attending school[ATSCH]).

The main results on fitting the three models to the SLID data are given in Table 2. For this analysis we did not distinguish single spells from the first spells of multiple-spell individuals. The standard errors of the estimated model coefficients needed for testing significance of each coefficient are obtained using the robust Binder (1992) approach. The design-based variance method used was Taylor linearization, assuming a survey design that is stratified with with-replacement selection of psu’s at the first stage. Coefficients found significant at the 5% level are given in bold.

Model 1 is conditional on the spell order and reduces to four models fitted separately. Visual observation of the differences in the estimated coefficients across spell orders and the significance of the model coefficients within a spell order were used to identify whether or not there appears to be a differential effect of the covariates on spell duration. For example, variables having significant negative effect for one spell order and not being significant for others are education levels lower then ‘M’, having a part time job, and attending school during the lay-off. All of them are significant for the first three spell orders and not for the fourth. This can be at least partly attributed to the small sample size for the fourth spells. Another example is the small company size which has an insignificant negative effect for the first two spell orders and has a significant positive effect for the fourth order. Age has significant negative effect for all four orders; however, the magnitude of effect decreases with the spell order. Thus, an increase in age by a year reduces the hazard of exit from an unemployment spell by about 5% for the first spell, about 3% for the second and the third spell, and 2% for the fourth. The estimated empirical cumulative baseline hazard functions for Model 1 are given in Figure 1. These functions are estimated as

$$\hat{H}_{0j}(t) = -\log \hat{S}_{0j}(t), \text{ for } j\text{=1 to 4.}$$

where $\hat{S}_{0j}(t)$ is the estimated baseline survivor function implicitly given by

$$\hat{S}_j(T_{ij}, x_{ij}) = \{\hat{S}_{0j}(T_{ij})\}^{\exp[x_{ij}^{\prime}(T_{ij})\hat{\beta}_j]}$$

and $\hat{S}_j(t, x)$ is the estimated survivor function (see Kalbfleisch and Prentice, 1980, page 84). From Figure 1 we can see that for durations up to 50 weeks the cumulative baseline hazards are ordered according to the spell order, with the dominance of the first order spell function. Also in this interval all the functions have a concave shape, essentially meaning that there is a positive time dependence of the exit rate (the longer the spell the higher the probability of exit) with the strongest dependence for the first order spells. For durations longer than 50 weeks the shape becomes convex suggesting negative time dependance for the longer spells with the strongest dependence for the third order spells.

Model 2 has the same beta coefficients for all spell orders. Numerically the estimated coefficient values are mostly situated between the estimates for the first and second order spells obtained for Model 1 due to the method of estimation and the sample shares that correspond to these orders. Since the entire sample is used to estimate the single set of coefficients, the estimates are more precise; however, this caused little change in which variables were found to have significant coefficients. The cumulative baseline hazard functions for this model are presented in Figure 2. The shape of the curves remains as in the Model 1 indicating the same type of time dependance for the exit rates. However, the ordering of the curves changed completely, with the dominance of spells of fourth order.

Model 3 is a single spell model resulting in 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. This fact could indicate that if Model 2 is approximately true, then miss-specification of the effect of spell order as is done in Model 1 still leads to reasonable estimates of the effects of risk factors. The cumulative baseline hazard function for Model 3 is given in Figure 3. Evidently its shape is the same as the shape of the hazard curves discussed previously.

The general conclusion from results given in Table 2 is that from the particular population of spells being studied, there is an interaction between the spell order and the covariates chosen for the model, as indicated by the different coefficients for the different orders of Model 1. However, if you were willing to restrict yourself to a model with the same coefficients for all spell orders (i.e. Model 2 or 3), the estimates of the covariate effects are approximately the same for the marginal model (Model 2) and the 'single spell' model (Model 3), suggesting that the 'single spell' model would give adequate results under this restriction, as compared to the extra effort needed to fit the marginal model.

We also examined the coefficient standard errors for Models 1 and 2 estimated by (i) the 'naive' method where the coefficient estimates are obtained under the assumption of independence of individuals and of the spells and then the 'naive' variance estimate is corrected for the spell dependency (which is a modification of the method described in Lin and Wei (1989) to include sampling weights); and by (ii) the robust Binder approach (which again estimates the coefficients by the 'naive' approach but corrects for the correlation between and within individuals induced by the survey design in the variance estimation). The standard errors obtained by correcting for the sample design effect are larger than the standard errors obtained by correcting only for the spell dependencies within individuals. The design-based variance estimation method used automatically accounts for the correlations between individuals and the correlation between multiple spells experienced by an individual because of the assumption of with-replacement sampling of psu's and because both individuals and spells are nested within psu's. (Each spell from the same individual is given the individual weight, since, given that an individual is chosen for the sample, the individual's spells are chosen with probability 1.) The ratio of the two standard errors could be considered as a measure of the magnitude of the sample clustering effect. The sample clustering effect on β for covariate SEX, for example, ranges between 1.35 and 1.65 for Model 1, while for Model 2 it is 1.80.

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 variance-corrected approach and a design-based approach. The first approach estimates the model parameters assuming the independence of the spells, and then corrects the naive covariance matrix to account for the within-individual dependencies. This approach does not account for the clustering between individuals induced by the sample design. The second approach defines the model parameters as finite population parameters. These parameters are then estimated accounting for possible unequal selection probabilities of individuals. A design-based variance estimation method that appropriately accounts for the correlations between individuals 'automatically' accounts for the unspecified dependancies of spells within individuals. For large sample sizes this design-based inference extends directly to the super-population which hypothetically underlies the finite population. The deficiency of the first approach is that it totally ignores the clustering between individuals. A possible disadvantage of the second approach is that it relies on the assumption of with replacement sampling of clusters of individuals. The two approaches coincide in the case of simple random sampling. We applied these approaches to three models: two of which use information on the spell order to specify the interaction of the spell order and covariates, and to allow for differential unspecified baseline hazards. The third model was a simple single spell model. We found that using information on the spell order affects the modelling of multiple spells.

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*, Vol.17, No. 4, 432-448

Chambless, L.E. and Boyle, K.E. (1985) Maximum Likelihood Methods for Complex Sample Data: Logistic Regression and Discrete Proportional Hazards Models. *Communications in Statistics - Theory and Methods*, 14(6), 1377-1392.

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. (1976). *Sampling Techniques*. Third edition. Wiley, New York.

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

Hamerle, A. (1989) Multiple-spell Regression Models for Duration Data. *Applied Statistics*, 38, No.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.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The statistical analysis of failure data*. John Wiley and Sons.

Keiding, N., Andersen, P.K. and Klein, J.P. (1997) The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in medicine* 16, 214-224.

Klein, J.P. (1992). Semiparametric Estimation of Random Effects Using the Cox Model Based on the EM algorithm. *Biometrics*, 48, 795-806.

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. and Wei, L.J. (1989) The robust Inference for the Cox Proportional Hazards Model. *Journal of American Statistical Association*, 84, 1074-1078.

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

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

Prentice, R.L. and Gloeckler, L.A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, 34, 57-67.

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 Kovacevic, 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.

Skinner, C. J. (1989). Domain Means, Regression and Multivariate Analysis. In *Analysis of Complex Surveys* (ed. By Skinner, C.J., Holt, D. and Smith, T.M.F.) Wiley, Chichester.

Therneau, T. M (1996). *Extending the Cox Model*. Technical Report No. 58. Section of Biostatistics. Mayo Clinic, Rochester, Minnesota.

Williams, R.L. (2000) A Note on Robust Variance Estimation for Cluster-Correlated Data. *Biometrics*, 56, 645-646.

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	**4141**	**2221**	**1920**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**
2 spells	**1915**	**1915**	**-**	**1154**	**761**	**-**	**-**	**-**	**-**	**-**	**-**
3 spells	**1044**	**1044**	**-**	**1044**	**-**	**612**	**432**	**-**	**-**	**-**	**-**
4 spells	**629**	**629**	**-**	**629**	**-**	**629**	**-**	**348**	**281**	**-**	**-**
5+ spells	**672**	**672**	**-**	**672**	**-**	**672**	**-**	**672**	**-**	**1158**	**415**
Total	**8401**	**6481**	**1920**	**3499**	**761**	**1913**	**432**	**1020**	**281**	**1158**	**415**

Table 2. *Estimated* β *coefficients for three models*

	Model 1				Model 2	Model 3
	Order1	Order2	Order3	Order4		
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	8386	4255	2345	1300	16286	16286
Censored	1913	759	432	281	3385	3385
Completed	6473	3496	1913	1019	12901	12901

The values significant at 5% level are bold

Figure 1. Cumulative Baseline Hazard — Model 1

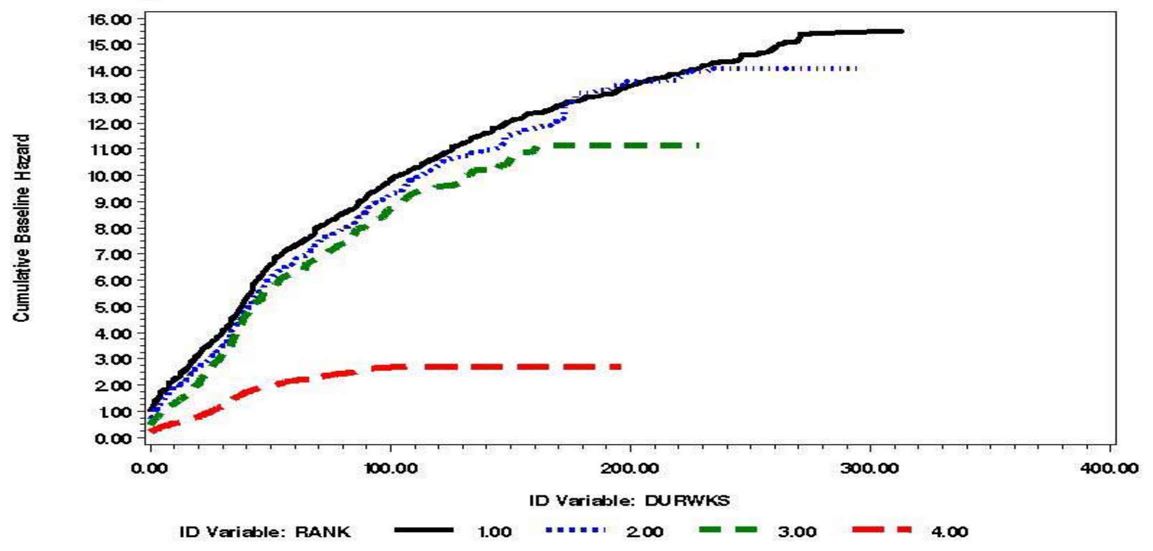


Figure 2. Cumulative Baseline Hazard — Model 2

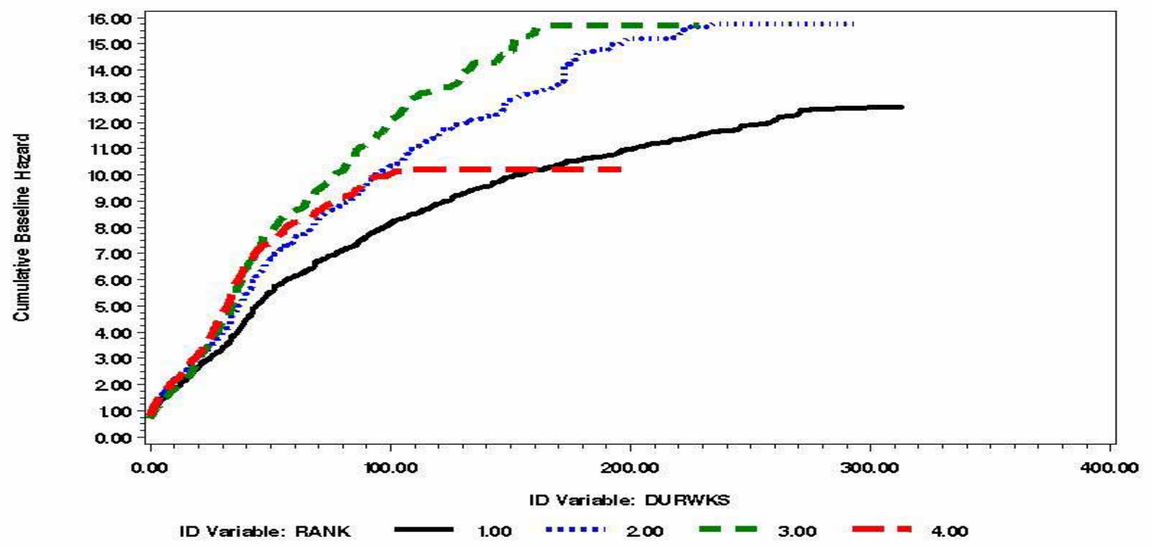
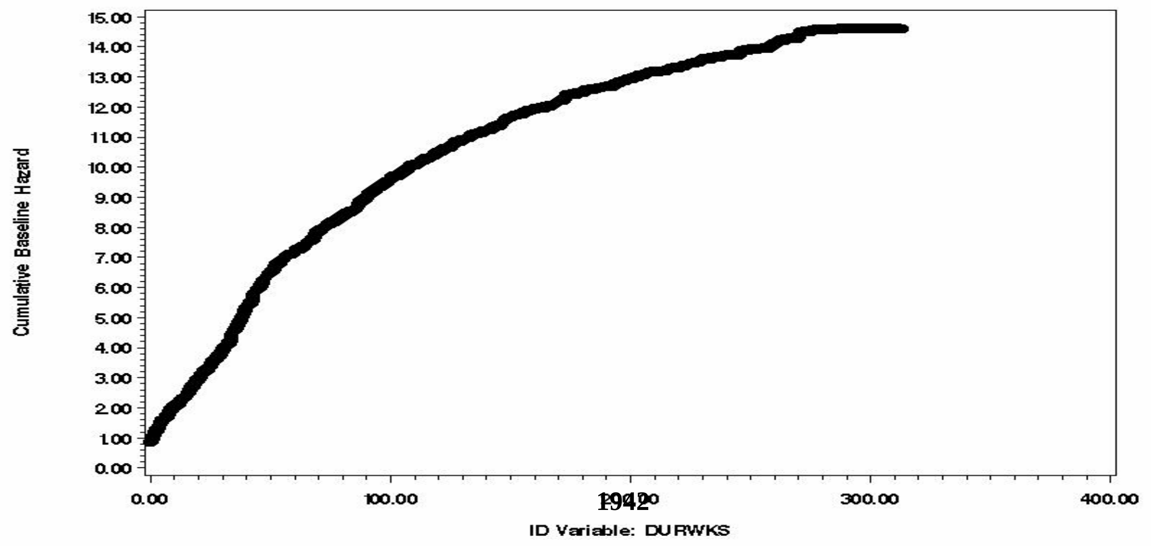


Figure 3. Cumulative Baseline Hazard — Model 3


Cumulative Baseline Hazard
ID Variable: DURWKS