Proceedings of Statistics Canada Symposium 2002
Modelling Survey Data for Social and Economic Research

MODELLING AND ANALYSIS OF DURATION DATA FROM LONGITUDINAL SURVEYS

J.F. Lawless and C. Boudreau[1]

ABSTRACT

Data from longitudinal surveys are used to help understand life history processes such as education, employment, fertility, and health. One aspect of this concerns the duration of spells or sojourns that individuals spend in certain states. This paper discusses methods for analyzing duration data that address features of longitudinal surveys: the use of complex sampling designs; intermittent follow-up of subjects; missing and mismeasured data; and losses to follow-up. The analysis of jobless spells in Statistics Canada's Survey of Labour and Income Dynamics is used for illustration.

1. INTRODUCTION

There is much interest in using data from longitudinal surveys to help understand life history processes such as education, employment, fertility, marriage, and health. One aspect of this concerns the durations of spells or sojourns that individuals spend in certain states, such as unemployment or marriage. Duration analysis and, more generally, life or event history analysis, are well developed areas with many standard methods (e.g., Andersen et al. 1993, Kalbfleisch and Prentice 2002, Lawless 2002). However, data collected through longitudinal surveys bring a variety of problems that standard methods do not fully handle. Our objective is to discuss such problems and review approaches for dealing with them.

Populations on which longitudinal surveys are based tend to be very heterogeneous, and individuals are typically sampled according to a complex survey design involving stratification and clustering. Even if analysis is directed at a specific subpopulation or process, this creates difficulties, as we discuss below. The way in which longitudinal data are collected in the survey can also cause problems: intermittent (e.g. annual) follow-up of individuals may lead to missing or mismeasured data and to difficulty in defining censoring times for persons lost to follow-up; loss to follow-up processes may not be ignorable; short duration follow-up provides limited information on life history processes that evolve over a long period of time; time-varying factors that affect a process are often not measured.

This paper is concerned with the use of longitudinal surveys to understand the life history processes of individuals. The conceptual framework used for modelling and analysis is the process that generates the individuals and their histories. Section 2 reviews methods of inference for duration data within this framework, stressing the need to deal with the survey design. Section 3 considers missing or mismeasured data involving initial conditions, losses to follow-up, the timing of duration-related events, and covariates. Section 4 discusses these issues in the context of unemployment spells in the Survey of Labour and Income Dynamics (SLID). Section 5 makes some concluding remarks.

2. DURATION ANALYSIS WITH SURVEY DATA

Suppose that a duration variable Y (sometimes called a lifetime or survival time) and a vector of covariates x are associated with individuals in some population and that the marginal distribution of Y given x is the focus of attention. Many duration problems deal with the length of time that an individual spends in some state and it will be

[1] University of Waterloo

convenient here to identify Y with the duration of a sojourn in a state denoted as *A*. Some degree of association between the *Y*'s, given the corresponding *x*'s, for individuals belonging to certain groups or clusters within the population is generally expected. This will be accounted for in the analysis below, but the association will not be explicitly modelled.

Suppose that the distribution of Y given x has probability density function $f(y|x;\theta)$. If one had a simple random sample S of individuals from the population then inferences about the model parameters θ could be based on the pseudo likelihood function

$$L(\theta) = \prod_{i \in s} f(y_i \mid x_i; \theta), \tag{2.1}$$

assuming temporarily that the durations y_i are all fully observed. However, the survey population, consisting of N individuals, is typically divided into strata $U_1, U_2, ..., U_H$ that consist of clusters of individuals forming primary sampling units (PSU's), and the survey sample S is chosen by randomly selecting PSU's within each stratum and then randomly sampling individuals from each PSU. Let $R_i = I(i \in S)$ denote whether individual i ($i = 1, ..., N$) is in the sample, and let the vector z_i denote design-related factors such as stratum and cluster information, so that the sample design inclusion probabilities π_i satisfy

$$\pi_i = \Pr(R_i = 1 \mid y_i, x_i, z_i) = \Pr(R_i = 1 \mid z_i) \qquad i = 1,..., N\ . \tag{2.2}$$

Note that x_i and z_i may share some components. If Y_i and Z_i are independent, given x_i, then $\Pr(y_i \mid x_i, R_i = 1) = f(y_i \mid x_i)$ and estimation of θ can be based on (2.1) or the corresponding pseudo score estimating function $U(\theta) = \partial \log L(\theta) / \partial\theta$, or

$$U(\theta) = \sum_{i=1}^{N} R_i \frac{\partial \log f(y_i \mid x_i; \theta)}{\partial \theta} = \sum_{i=1}^{N} R_i U_i(\theta). \tag{2.3}$$

We term $U(\theta)$ a pseudo score function because the Y_i ($i \in S$) are not in general independent, so *(2.1)* is not their joint probability density function. However, solving $U(\theta) = 0$ provides a consistent estimator $\hat{\theta}$ of θ under mild conditions.

If $Pr(y_i|x_i,z_i) \neq Pr(y_i|x_i)$ then $Pr(y_i|x_i,R_i = 1) \neq f(y_i|x_i)$, and (2.3) in general yields inconsistent estimates of θ. In this case a weighted pseudo score function

$$U_W(\theta) = \sum_{i=1}^{N} \frac{R_i}{\pi_i} \frac{\partial \log f(y_i \mid x_i; \theta)}{\partial \theta} = \sum_{i \in s} w_i U_i(\theta) \tag{2.4}$$

can be used for estimation (e.g. Pfefferman *1993;* Thompson *1997*, Sec. 6.2). Solving $U_W(\theta) = 0$ yields an estimator $\hat{\theta}$ that is consistent for θ, though it has larger variance than estimators based on (2.3) when both are consistent.

Pfefferman *(1993)* provides an insightful review of weighted versus unweighted estimation and of the potential uses of the model $f(y|x;\theta)$ when Y is not independent of design factors Z, given x. We suppose for most of the discussion here that the covariates x include sufficient information that the survey design is ignorable and (2.3) can be used. However, the methodology is easily extended to include weights, as in (2.4).

2.1 Estimation for Parametric Duration Distributions

Suppose that Y given x has survivor function $S(y \mid x; \theta) = Pr(Y \geq y \mid x; \theta)$ and hazard function $\lambda(y|x;\theta) = f(y|x;\theta)/S(y|x;\theta)$. Suppose also that individual i is followed over the calendar time period (L_i, R_i) and consider observation of a sojourn in state A that covers the time period $(t_i, t_i + \tilde{y}_i)$, which overlaps (L_i, R_i). If $t_i \leq L_i$ then the spell duration $\tilde{y}_i$ satisfies $\tilde{y}_i \geq L_i - t_i$ and is left-truncated at $v_i = L_i - t_i \geq 0$. *If* $t_i + \tilde{y}_i > R_i$ then we observe only that $\tilde{y}_i > R_i - t_i$ and $\tilde{y}_i$ is

right-censored at $c_i = Ri - t_i$. With the notation $y_i = \min(\tilde{y}_i, c_i)$ and $\delta_i = I(y_i = \tilde{y}_i)$, the pseudo score component $U_i(\theta)$ in (2.3) is then replaced by (e.g., Lawless *2003,* Sec. 5.2)

$$U_i(\theta) = \delta_i \frac{\partial \log \lambda(y_i \mid x_i; \theta)}{\partial \theta} - \frac{\partial}{\partial \theta} \int_{v_i}^{y_i} \lambda(t \mid x_i; \theta) dt \,. \tag{2.5}$$

Based on the sample design described earlier, the pseudo score (2.3) with components (2.5) can be written as

$$U(\theta) = \sum_{h=1}^{H} \sum_{c=1}^{C_h} \sum_{i \in s_{hc}} U_i(\theta), \tag{2.6}$$

where C_h is the number of clusters or PSU's taken from stratum h and S_{hc} denotes the individuals sampled from cluster c in stratum h. We assume that the duration times of individuals in distinct clusters are independent. Under mild conditions the estimator $\hat{\theta}$ obtained by solving $U(\theta) = 0$ is then asymptotically normal with mean θ and covariance matrix that is estimated consistently by

$$V(\theta) = I(\hat{\theta})^{-1} \hat{Var}(U(\theta)) I(\hat{\theta})^{-1}, \tag{2.7}$$

where $I(\theta) = -\partial U(\theta)/\partial \theta'$ and

$$\hat{Var}(U(\theta)) = \sum_{h=1}^{H} \sum_{c=1}^{C_h} \left(\sum_{i \in s_{hc}} U_i(\hat{\theta}) \right) \left(\sum_{i \in S_{hc}} U_i(\hat{\theta}) \right)' . \tag{2.8}$$

This allows for association among the duration times of individuals in the same cluster. Existing software for accelerated failure time or other parametric models can be used to obtain $\hat{\theta}$ and $I(\hat{\theta})$;a little extra programming is generally necessary to compute (2.8).

2.2 Semiparametric Cox Models

The Cox (1972) proportional or multiplicative hazards model is widely used. It takes the hazard function for Y given x to be of the form

$$\lambda(y \mid x) = \lambda_0(y) \exp(\beta' x) \tag{2.9}$$

where $\lambda_0(y)$ is an arbitrary "baseline" hazard function and β is a vector of regression coefficients. Well known partial likelihood methods provide estimates of β and nonparametric estimates of $\Lambda_0(y) = \int_0^y \lambda_0(t) dt$ under simple random sampling, and observation schemes allowing left-truncation or right-censoring (e.g. Kalbfleisch and Prentice 2002, Ch. 4).

The standard methods can be extended to handle within-cluster association in duration times (e.g. Lee et al. 1992). Spiekerman and Lin (1998) and Boudreau and Lawless (2001) have shown how to deal with both cluster and stratum effects, and time-varying covariates. For example, in some settings the stratified Cox model with hazard functions

$$\lambda(t \mid x_i, i \in U_h) = \lambda_{0h}(t) \exp(\beta' x_i(t)) \tag{2.10}$$

is useful. Estimation of β and of the cumulative hazard functions $\Lambda_{0h}(t)$ can be based on the standard estimating functions from the case of simple random sampling (Kalbfleisch and Prentice 2002, Sec. 4.4). Boudreau and Lawless (2001) derive variance estimates analogous to (2.7) and demonstrate that they are easily obtained using the

S-Plus software function coxph, with the cluster and strata options. Their methodology can also handle weights. An illustration is provided in Section 4.

We remark that Binder (1992) and Lin (2000) have discussed weighted estimation for the Cox model. They both considered design-based variance estimation, and Lin also discussed model-based estimation as done here. Lin allows more general survey designs than we discuss but ignores association in duration times at the population level. The variance estimates given by our methods (with (2.8) - (2.9) extended to include weights) and his (using software such as SUDAAN for the design-based portion of the variance) are very close for settings that we have examined.

3. MISSING OR MISMEASURED DATA

3.1 Initial Conditions and Left-Truncation

An individual is typically followed over some calendar period (L_i, R_i) in a longitudinal survey, but analysis may require information about the individual prior to time L_i. One such instance occurs when an individual is already in state A at time L_i, with interest focused on the duration Y of sojourns in that state. If the time $t_i \le L_i$ at which an individual entered the state is known then, provided t_i and Y_i are independent, the relevant distribution for the sojourn duration is the left-truncated density $f(y)/S(vi)$, where $v_i = L_i - t_i$. This gives the pseudo likelihood score function (2.5). In some cases the distribution of Y_i may depend on the time t_i of entry into state A, in which case $f(y)$ and $S(y)$ are replaced with $f(y|t)$ and $S(y|t)$. If, however, Y_i and t_i both depend on unmeasured factors then (2.5) is no longer appropriate (e.g. Keiding 1992, Lawless and Fong 1999). The use of design weights will not alleviate this. Such "initial conditions" problems have received considerable discussion in economics and the social sciences (e.g. Heckman and Singer 1985, Hoem 1985).

The entry time t_i into the state A may also be unknown or measured with error. In the former case, let $g(v)$ denote the p.d.f. of $V_i = L_i - t_i$, given that an individual is in state A at time L_i and assuming independence of Y_i and t_i. For simplicity, dependence on covariates is suppressed in the notation. If t_i, and hence v_i, are unobserved, the appropriate p.d.f. for the observed data is

$$f^*(w_i) = \int_0^\infty g(v)\frac{f(v+w_i)}{S(v)}dv\,, \tag{3.1}$$

where $w_i = t_i + y_i - L_i$ is the duration of the remaining sojourn in state A after time L_i. Either w_i or a censoring time for it are assumed to be observed.

A model for $g(v)$ is thus required when t_i is unobserved. This can sometimes be specified through knowledge of the flow rate into state A over time. Sometimes an argument is made for assuming that $g(v) = S(v)/\mu$, where $\mu = E(Y)$ is the average duration in state A; this corresponds to a flow rate into state A that is constant over time and can also be obtained through equilibrium arguments in a renewal process. If $g(u) = S(v)/\mu$ then (3.1) reduces to $S(wi)/\mu$, which is the well known forward recurrence time distribution in renewal theory (e.g., Ross 1983). Note that it may be possible to test a hypothesized model for g(v). In particular, if there are individuals whose sojourns Y_i in state A are untruncated then an estimate $\hat{S}(y)$ can be obtained from them. The empirical distribution of the w_i's for left-truncated sojourns can then be compared with $\hat{S}(y)/\hat{\mu}$,or, more generally, the distribution implied by (3.1).

In the absence of a suitable model for $g(v)$, an acceptable procedure is simply to discard left-truncated y_i's, using only sojourns in state A that start while an individual is under follow-up (e.g., Aalen and Husebye 1991). Boudreau and Lawless (2002) examine the loss of information in doing this when (3.1) is actually available. A practical issue is that in some settings most observations may be subject to some degree of left-truncation.

If the value of t_i is recorded with error of measurement due to recall or other factors, one approach is to model the measurement error along with the joint distribution $g_1(t_i)\, f(y_i) / S(v_i)$ of t_i and Y_i to get a pseudo likelihood based on the observed data. This can be very difficult, and the possible dependence of the measurement error on v must be

considered. Rather than incorporate formal modelling into the estimation process, one might carry out a sensitivity analysis to assess the effect of mismeasurement on standard estimates. Lawless (2003, Sec. 8) provides an illustration involving the recall of the date at which breast feeding of infants terminated.

3.2 Effects of Intermittent Observation

When panel members are interviewed at rather widely spaced time points, the accuracy of event times and other variables over the period since the preceding interview should be considered. Seam effects can occur, in which individuals tend to locate events near times of interviews (e.g., Kalton and Miller 1991). Problems involving missing or mismeasured covariates are very difficult to handle with duration data, especially when the censoring process is related to the covariates (e.g., Kalbfleisch and Prentice 2002, Ch. 11). Modelling of the covariate and censoring processes is unavoidable in many cases. Common sampling approaches such as imputation have not been carefully examined, but would also require strong assumptions about censoring. This area is in need of development.

Missing information about event times or durations due to intermittent observation can be more readily handled, though supplementary modelling may be needed. For example, if entry into state A has occurred at an unknown time between the preceding and current interview dates, then a model for the time of entry is needed, as discussed in Section 3.1.

Intermittent observation can also affect the validity of standard censoring assumptions. For example, if an individual is in the midst of a sojourn in state A at the year t interview and then becomes lost to follow-up by the year t + 1 interview time, then it is standard practice to assign a censoring time based on the current duration in state A at the year t interview. However, this requires that the loss to follow-up mechanism is independent of the individual's duration process over the period following the interview, given covariates measured up to year t. This assumption is likely violated in many settings. The only clear resolution of this problem is to trace some individuals lost to follow-up (e.g. Farewell et al. 2002), but sensitivity analysis can be used to assess the effect of an hypothesized dependent loss-to-follow-up mechanism.

4. ILLUSTRATION: UNEMPLOYMENT SPELLS IN SLID

We provide here a brief illustration of issues based on Statistics Canada's Survey of Labour and Income Dynamics (SLID). We consider data from the first panel, for which follow-up started in January 1993 and continued for six years (barring loss to follow-up). Panel selection is based on a stratified multistage sampling design in which the PSU's are groups of households forming enumeration areas (EA's). Sampling involves a random selection of households within sampled PSU's; all individuals over 16 years of age in a household are followed longitudinally in the survey. Panel members are interviewed annually in January, at which time information on events that occurred over the previous calendar year is collected. Detailed information on SLID is available from its web site, which can be accessed from http://www.statcan.ca.

The example given here deals with the duration of jobless spells for persons who are in the labour force and don't suffer from long term disabilities. We discuss issues relating to missing or mismeasured data below, but first illustrate a semiparametric proportional hazards (PH) analysis of the duration of spells that started in the calendar year 1993. These data involve a total of 2313 observed spells, 104 of which were right-censored.

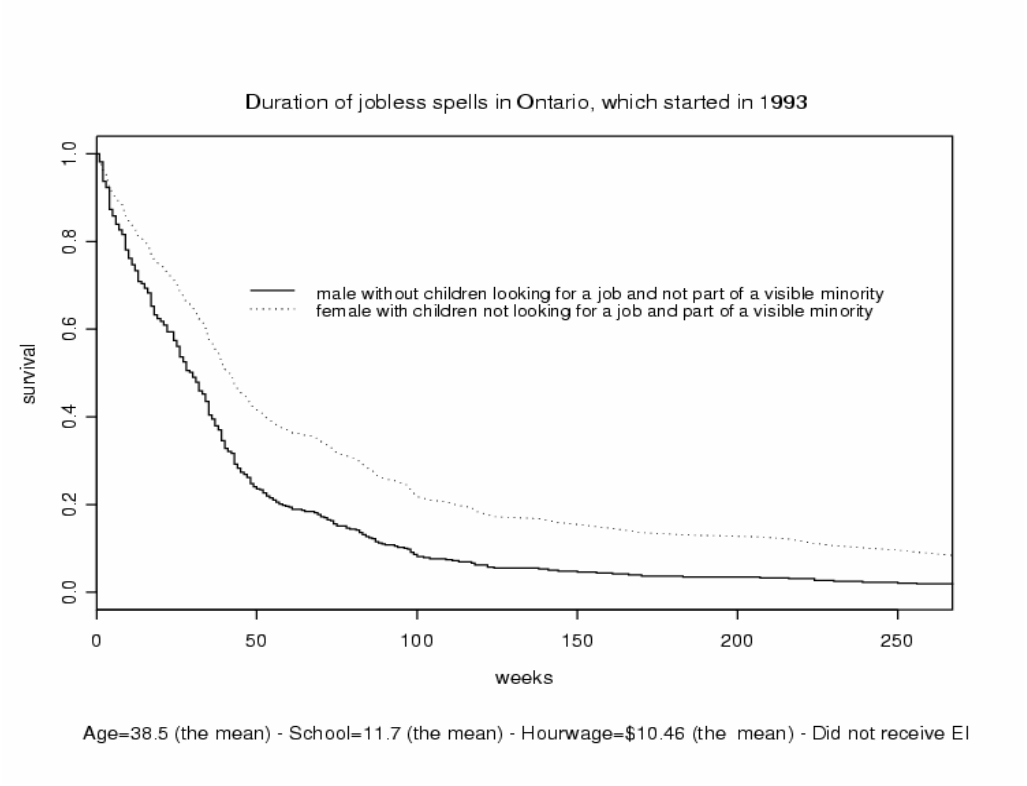
Table 1 shows results for a model involving nine covariates: SEX (= 1 if female, 0 if male); LOOKED (= 1 if looking for work, 0 if not); CHILDREN (= 1 if individual has one or more children, 0 otherwise); AGE (= age as of January 1993); SCHOOL (= number of years of schooling); HOURWAGE (= wage in dollars per hour); EI (= 1 if person received employment insurance, 0 if not); MINORITY (= 1 if person is a member of a visible minority group, 0 if not); WINTER (= 1 if person lost previous job during months November - April). Two-factor interactions involving these covariates are also considered. The model fitted was a stratified Cox PH model (Kalbfleisch and Prentice 2002, Section 4.4) with provinces forming the strata. Estimates and standard errors based on both weighted and unweighted estimating functions (Boudreau and Lawless 2001) are shown in Table 1. The weighted estimates used individual sampling weights provided by SLID that account for both sample design and a loss to follow-up adjustment. Two additional sets of standard errors are shown for them: the finite sample design-based standard

errors of Binder (1992) and the analogous standard errors of Lin (2000), which are based on Binder's variance estimates plus an additional term for the superpopulation process. Our standard errors for both weighted and unweighted estimates are based on using households as clusters. The use of EA's as clusters produced standard errors that were very close to these. For the unweighted case, "independence" (and technically incorrect) standard errors are shown; they are based on treating the unweighted analysis as though it arose from a sample of mutually independent duration times.

The three standard errors for the weighted analyses are virtually identical, indicating that the superpopulation correction term of Lin (2000) adds only a tiny amount to Binder's (1992) designbased standard errors, and also that our simple robust variance estimation procedures work well here. The unweighted and weighted estimation procedures give estimates that generally agree well; the largest difference is for the LOOKED covariate, where they differ by just under two (weighted) standard errors. The independence standard errors for the unweighted analysis are not very different form the robust standard errors, indicating only a small effect due to association of jobless durations for persons in the same household.

Diagnostic checks on the model based on the unweighted analysis (e.g. Lawless 2002, Section 7.2) indicate only one significant departure from assumptions: the effect of the WINTER covariate decreases with time. This suggests that losing one's job during the winter months is associated with an increasing chance of finding employment over the next 12 months, but does not much affect the hazard function for return to employment after about 12 jobless months.

Figures 1 and 2 show illustrative plots of survivor functions for the duration of a jobless spell.



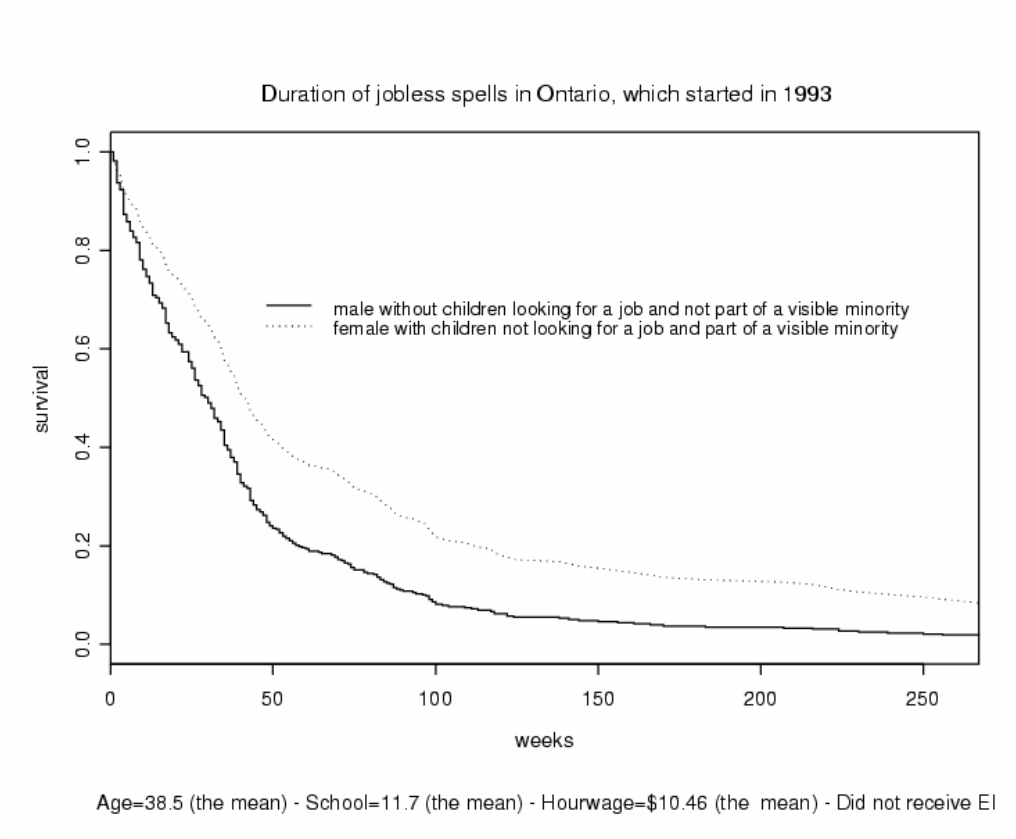


Table 1. Estimates and Standard Errors of PH Covariate Effects for Jobless Spells

Covariate	Unweighted estimate	SE	SE (Ind.)	Weighted estimate	SE	SE (Lin)	SE (Binder)
SEX	-0.561	0.224	0.212	-0.628	0.294	0.295	0.295
LOOKED	-0.119	0.051	0.056	0.023	0.075	0.075	0.075
CHILDREN	-0.714	0.132	0.128	-0.847	0.196	0.196	0.196
AGE	-0.025	0.006	0.005	-0.031	0.009	0.009	0.009
SCHOOL	0.001	0.009	0.010	-0.001	0.012	0.012	0.012
HOURWAGE	0.062	0.013	0.012	0.037	0.021	0.021	0.021
EI	-0.368	0.130	0.125	-0.216	0.178	0.178	0.178
MINORITY	-0.183	0.204	0.233	-0.098	0.244	0.244	0.244
WINTER	0.205	0.092	0.080	0.242	0.135	0.135	0.135
SEX*SCHOOL	0.033	0.015	0.015	0.036	0.020	0.020	0.020
SEX*CHILDREN	-0.007	0.088	0.089	-0.015	0.122	0.122	0.122
AGE*CHILDREN	0.023	0.004	0.004	0.025	0.006	0.006	0.006
SEX*HOURWAGE	-0.022	0.008	0.008	-0.008	0.012	0.012	0.012
WINTER*LOOKED	-0.100	0.105	0.097	-0.170	0.150	0.150	0.150
AGE*HOURWAGE	-0.001	0.0004	0.0003	-0.001	0.0005	0.0005	0.0005
AGE*EI	0.002	0.004	0.004	0.015	0.006	0.006	0.006

Many issues related to points discussed in Section 3 arise with SLID data. Left-truncation of duration spells occurs for persons in the panel who were unemployed as of January 1, 1993. However, information on the start of these spells is not available from the survey, and so we chose here to base estimation only on spells that began after January 1, 1993. In order to use the left-truncated spells as well, we would have to model the flow into the unemployed state for persons becoming unemployed prior to 1993 and remaining unemployed on January 1, 1993.

Losses to follow-up, or attrition, are common in SLID. For example, the initial 1993 panel comprised about 15,000 households and 31,000 persons age 16 or over. The percentages of these persons that were subsequently interviewed in January of the years 1994, 1995, and 1996 were 89.6%, 86.5% and 85.2%, respectively. The impact of the losses to follow-up on the estimation of the duration distribution is greatest for long durations. For example, in Figure 1 it is seen that some jobless spells have very long durations; the estimation of this part of the distribution could, however, be heavily influenced by any non-ignorable patterns in panel attrition.

Various issues arise in the measurement of job duration, including the effects of respondent recall and how the starts and ends of spells of employment are recorded. In SLID the latter are recorded by week of occurrence, so there is only a small grouping effect for duration times. Other factors, including seam effects, also come into play (e.g. see Cotton and Giles 1998).

The issue of weights in surveys such as SLID is rather complicated. “Longitudinal” weights can be calculated for each person in the initial 1993 panel, but as has been indicated earlier, weights for duration analysis need to be adjusted over time, to deal with losses to follow-up, if we want protection against the effects of non-ignorable sampling and attrition. A spell that began in 1993 and ended in 1994, for example, would have the start date reported in January 1994 and the end date reported in January 1995. For simplicity of analysis one would presumably use either the January 1994 or January 1995 weights, but neither choice is fully satisfactory in all circumstances. It should also be noted that individuals who join households in the original panel are interviewed in subsequent years. They have initial longitudinal weights of zero but can be included in analyses, such as those discussed in this paper.

5. CONCLUDING REMARKS

Standard methods of survival or duration analysis can be adapted to data from longitudinal surveys by utilizing methods of robust variance estimation and, if necessary, weights to counter the effects of non-ignorable sampling or losses to follow-up. Weights must in some cases be time-varying, however, and methods of choosing them deserve further study. There are connections here with the work of Robins and others (e.g., Robins and Rotnitzky 1995).

We note a few other topics that deserve study. First, based on a desire to develop models that describe the life history processes of individuals, there is a need for diagnostic checking of duration models. Methods exist for standard unweighted analysis but the effects of cluster sampling have received little study. Methods based on comparisons of weighted and unweighted estimates (e.g., Pfefferman 1993) could also be developed.

The study of repeated spells, or transitions among states, for an individual is of much interest, and raises the issue of multivariate duration models and measures of association. Multilevel modelling provides one approach (e.g., Bandeen-Roche and Liang 1996) but it seems desirable to develop simpler methods that model individual level processes but use robust variance estimation techniques to deal with higher level cluster effects.

Empirical study and reasonably simple methods of dealing with measurement error, missing covariate values, and possibly non-ignorable losses to follow-up deserve attention, as mentioned previously.

Finally, the life history processes associated with individuals very often depend substantially on time-varying factors. Methods of duration analysis that incorporate time-varying covariates are available, but there nonetheless are issues concerning how best to quantity or describe specific life history processes, in order to gain a fuller understanding of their evolution.

ACKNOWLEDGEMENTS

This research was supported in part by a grant to the first author from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Aalen, O. and Husebye, E. (1991). Statistical analysis of repeated events forming renewal processes. *Statistics in Medicine* ***10,*** 1227-40.

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Bandeen-Roche, K.J. and Liang, K.-Y. (1996). Modeling failure-time associations in data with multiple levels of clustering. *Biometrika* ***83,*** 29-39.

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

Boudreau, C. and Lawless, J.F. (2001). Survival analysis based on Cox proportional hazards models and survey data. University of Waterloo Dept. of Statistics and Actuarial Science Working Paper 2001-10.

Boudreau, C. and Lawless, J.F. (2002). Efficiency and robustness issues involving missing or mismeasured left-truncation times in survival analysis. Manuscript.

Cotton, C. and Giles, P. (1998). The seam effect in the survey of labour and income dynamics. Statistics Canada Income and Labour Dynamics Working Paper Series, No. 98-18.

Farewell, V.T., Lawless, J.F., Gladman, D.D. and Urowitz, M.B. (2002). Analysis of the effect of lost-to-followup on the estimation of mortality from patient registry data. Submitted to *Applied Statistics*.

Heckman, J.J. and Singer, B. (1985). Social science duration analysis. In *Longitudinal Analysis of Labor Market Data*. Cambridge University Press, Ch. 2, eds. J.J. Heckman and B. Singer.

Hoem, J. (1985). Weighting, misclassification, and other issues in the analysis of survey samples of life histories, In *Longitudinal Analysis of Labor Market Data*. Cambridge University Press, Ch. 5, eds. J.J. Heckman and B. Singer.

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

Kalton, G. and Miller, M.E. (1991). The seam effect with social security income in the Survey of Income and Program Participation. J. *Official Statistics 7,* 235-245.

Keiding, N. (1992). Independent delayed entry. In *Survival Analysis: State of the Art,* 309-326. J.P. Klein and P.K. Goel, editors. Kluwer, Dordrecht.

Lawless, J.F. (2002). *Statistical Models and Methods for Lifetime Data,* 2nd edition. John Wiley and Sons, New York.

Lawless, J.F. (2003). Event history analysis and longitudinal surveys. To appear in *Analysis of Survey Data,* eds. R.L. Chambers and C.J. Skinner. John Wiley and Sons, Chichester.

Lawless, J.F. and Fong, D.Y.T. (1999). State duration models in clinical and observational studies. *Statist. Med.* ***18,*** 2365-2376.

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

Pfefferman, D. (1993). The role of sampling weights when modeling survey data. *Int. Statist. Rev.* ***61,*** 317-337.

Robins, J.M. and Rotnitzky, A. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika,* ***82,*** 805-820.

Ross, S.M. (1983). *Stochastic Processes*. John Wiley and Sons, New York.

Spiekerman, C.F. and Lin, D.Y. (1998). Marginal regression models for multivariate failure time data. *J. Amer. Statist. Assoc.* ***93,*** 1164-1175.

Thompson, M.E. (1997). *Theory of Sample Surveys*. Chapman and Hall, London.