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

INTERVAL CENSORING OF SMOKING CESSATION IN THE NATIONAL POPULATION HEALTH SURVEY

M. E. Thompson[1] and N. Pantoja Galicia[2]

ABSTRACT

The first three cycles of the National Population Health Survey allow the methodology of event history analysis to be applied to smoking initiation, cessation and relapse. One issue of interest is the relationship between smoking cessation and pregnancy. If a longitudinal respondent who is a smoker at the first cycle ceases smoking by the second cycle, we know the cessation time to within an interval of length at most a year, since the respondent is asked for the age at which she stopped smoking, and her date of birth is known. We know also whether she is pregnant at the time of the second cycle, and whether she has given birth since the time of the first cycle. For many such subjects, we know the date of conception to within a relatively small interval. If we knew the time of smoking cessation and pregnancy period exactly for each member who experienced one or both of these events between cycles, we could model their temporal relationship through their joint intensities. In this paper we consider the effect of the interval censoring of cessation time on intensity parameter estimation.

KEY WORDS: Longitudinal surveys; Temporal ordering; Survival analysis.

1. LONGITUDINAL SURVEYS AND EVENT SEQUENCES

It is rarely possible to establish causation in observational studies. At the same time, a causal interpretation of an association between events can sometimes be made more plausible if one of the events tends to precede the other closely in time. Longitudinal surveys allow us to observe, fully or partially, the sequences of events for individuals.

The National Population Health Survey (NPHS) is Statistics Canada's national longitudinal survey on health and health behaviour. Responses have now been collected from the longitudinal sample for four cycles. The first three, collected in 1994-95, 1996-97 and 1998-99, are available through the Research Data Centres. The sampling design for the initial cycle selected one individual at random from each of about 17,000 households across the country, and there are 14619 subjects who are accounted for in all of the first three cycles. These individuals form the longitudinal sample up to 1999. Thus we have a large number of subjects for whom it is possible to observe sequences of events as they are recalled by the subject at two year intervals.

In principle, the cycles of the National Population Health Survey allow the methodology of event history analysis to be applied to point events such as smoking initiation, cessation and relapse. At the same time, these events tend to be interval censored. For example, one issue of interest is the relationship between smoking and pregnancy. If a longitudinal respondent who is a smoker at the n-th cycle ceases smoking by the $(n+1)$-th cycle, we have a cessation time (assuming a correct response) to within an interval of at length at most a year, since the respondent is asked for the age at which she stopped smoking, and her date of birth is known. We know also whether she reports being pregnant at the time of the $(n+1)$-th cycle, and whether she has given birth since the time of the n-th cycle. For many such subjects, the file can be linked to a household file showing birth dates, and we can infer a relatively small interval for the date of becoming pregnant.
The purposes of this paper are twofold. First, we will discuss several notions of temporal association and ordering, and the concept of one type of event "triggering" another. Second, we will outline the construction of tests for these temporal relationships, showing how the tests may be applied in the context of smoking cessation in the NPHS.

[1] University of Waterloo, Waterloo ON, Canada N2L 3G1 (methomps@uwaterloo.ca)
[2] University of Waterloo, Waterloo ON, Canada N2L 3G1 (npantoja@uwaterloo.ca)

2. ASSOCIATION AND CAUSAL MODELLING

According to a summary by Cox (1992) of the writing of Bradford Hill (1937, 1965), there are circumstances under which an "effect" obtained in an observational study is relatively likely to have a causal interpretation: the effect

(a) is large

(b) is reproduced in independent studies

(c) shows a monotone relation with "dose"

(d) corresponds to a "natural experiment"

(e) behaves appropriately when the potential cause is applied, removed and then reinstated

(f) is consistent with subject matter knowledge or

(g) is predicted by reasonably well-established theory.

Not in the list is that the effect is consistent with subject attribution, probably because Bradford Hill was not thinking of causes of behaviour. In the NPHS example we do have some subject attribution, to decide whether or not to take account of.

Order is an important consideration in applying causal interpretations to associations. Cox (1992) discusses associated events C and E where C is a candidate cause of E:

> [To infer that C as a (candidate) cause of E is "spurious", through E being independent of C, given B] it is necessary to restrict the variables C, E and B, not least to express the essential asymmetry between cause and effect. The simplest is to insist that B occurs before C in time which in turn occurs before E. …Sometimes spatial proximity can be used as a basis for ordering effects instead of temporal ordering. A third possibility is that specific subject-matter knowledge is used to establish a presumed causal ordering of variables.

It is natural to ask whether longitudinal data, having built-in temporal order, are indeed especially valuable for inference. In attempting to sort out the relationships of variables, is it sometimes helpful to be able to verify a *tendency* for E to follow C, and a *tendency* for C to follow B?

To demonstrate that *ordering* is a type of *association*, in the next section we propose some characterizations of each. We do this for pairs of duration times T_1 and T_2, and for event sequences with events of two types.

3. TEMPORAL ASSOCIATION AND ORDERING

3.1 Joint Duration Time Distributions

We might say that duration times T_1 and T_2 are *weakly associated* if they are positively correlated.

Sufficient conditions for weak association to occur would be that $T_1 = Z + \varepsilon_1$, $T_2 = Z + \varepsilon_2$, with Z, ε_1 and ε_2 positive and independent, or more generally that the joint distribution of T_1 and T_2 is a suitable mixture of jointly independent distributions. In the first case particularly, it is natural to think of T_1 and T_2 as associated because both are preceded (strictly) by another duration time Z, a candidate for a causal event. In the second case, a similar mechanism where the ordering is not strict may be operating.

We might say that duration times T_1 and T_2 are *strongly associated* if knowing a lower bound for T_1 moves the mass for T_2 further along, i.e. if for each t_1, $t_2 > 0$ the conditional survivor functions satisfy

$$Pr(T_2 > t_2 \mid T_1 > t_1) > Pr(T_2 > t_2) \tag{1}$$

This is easily seen to be equivalent to

$$F(t_1, t_2) > F_1(t_1) F_2(t_2) \tag{2}$$

for all t_1, $t_2 > 0$, which can be shown in turn to imply weak association.

A sufficient condition for strong association would be that T_1 is exponential and $F_2(t_2 \mid T_1 = t_1)$ increases with t_1.

Temporal ordering for joint duration distributions is in a sense a kind of association, namely an association which is "one-sided".

The first ordering definition is one which is incontestably an order relationship. We say that T_1 is a *strict precursor* of T_2 if $T_2 - T_1$ is a non-negative random variable.

If T_1 is a strict precursor of T_2, T_1 and T_2 are weakly associated if T_1, $T_2 - T_1$ are independent, and strongly associated if T_1 are also exponential. A strict ordering is seldom met with in the kinds of applications we are talking about, and it is of interest to consider other definitions.

The notion of "weak precursor" can form the basis of non-parametric tests.

T_1 is a *weak precursor* of T_2 if $P(T_1 < T_2)$ under the joint distribution is greater than $P(T_1 < T_2)$ under the assumption of independence of T_1 and T_2.

Somewhat analogous to strong association for duration distributions is the "strong precursor" relationship.

T_1 is a *strong precursor* of T_2 if

$$F_2(t_1 \mid T_1 = t_1) > F_2(t_1) \tag{3}$$

for all t_1.

It is not difficult to show that a precursor in the strong sense is also a precursor in the weak sense. At the same time, it might not seem strong enough. A causal connection might be more plausible, the more closely T_2 tends to follow T_1. There is also a "close precursor" concept which neither implies nor is implied by any of the others, but which implies a local association of T_1 and T_2.

T_1 is a *close precursor* of T_2 if for positive numbers δ and $\kappa(t_1)$ we have

$$\frac{F_2(t_1 + \kappa(t_1) \mid T_1 = t_1)}{F_2(t_1 \mid T_1 = t_1)} < \frac{F_2(t_1 + \kappa(t_1))}{F(t_1)} - \delta \tag{4}$$

3.2 Joint Intensities

Temporal association for recurrent events which do not have strong logical connections is very naturally a *local* property, and will have both conditional and 'marginal' versions.

Marginal version:

Consider a point process with two kinds of events × and ∘. Suppose the two kinds of events have an unconditional joint intensity function $\lambda(t_1, t_2)$.

We could say the events × and ∘ are *temporally associated* if for each t_1, and each t_2 within some distance $\kappa(t_1)$ of t_1, we have

$$\lambda(t_1, t_2) > \lambda_1(t_1)\lambda_2(t_2). \tag{5}$$

Temporal order for recurrent events must involve closeness as well as order. In a 'conditional' version we could say event ∘ is a *close precursor* of event × if for suitably chosen $\kappa(s)$ and $\delta(s) > 0$ we have

$$\lambda_2(u \mid H(s-) \text{ and } \circ \text{ at } s) > \lambda_2(u \mid H(s-)) + \delta(s) \tag{6}$$

for u between s and $s + \kappa(s)$. Here $H(s-)$ represents the history of the joint process before time s.

It is possible, though not necessary, for two events ∘ and × to be precursors of each other in this sense.

4. TRIGGERING MODELS

Triggering is a special case of temporal ordering, where a causal model is explicit. It seems most natural to provide a conditional formulation for triggering, as follows.

4.1 Joint Duration Time Distributions

We might say that T_1 *triggers* T_2 if the occurrence of the end of duration T_1 changes the hazard function of T_2. In a simple example, suppose that if T_1 were infinity, T_2 would have survivor function $F_{02}(t_1)$ and hazard function $\lambda_{02}(u)$. However, if $T_1 = t_1$, then T_2 has hazard function $\lambda_{02}(u)$ before t_1 and $e^{\beta}\lambda_{02}(u)$ after t_1, where $\beta > 0$. It is possible to show that, unconditionally, T_2 then has survivor function

$$F_2(t) = F_{02}(t)\left[\int_0^t f_1(t_1)e(\beta, t_1, t)dt_1 + F_1(t)\right], \tag{7}$$

where

$$e(\beta, t_1, t) = exp\left\{-(e^{\beta} - 1)\int_{t_1}^{t} \lambda_{02}(u)du\right\}. \tag{8}$$

From this expression it is easy to show that T_1 is a strong precursor of T_2. With duration times it is possible to formulate this kind of long term triggering property. Local triggering can also possible be formulated, as a special case of the same notion for joint intensities below.

The "local" formulation in terms of a short term scale change in the hazard of T_2 leads to a density relation

$$f_2(t \mid T_1 = t_1) = e^{\beta} f_{02}(t) e(\beta, t_1, t) \tag{9}$$

for each t_1, and t between t_1 and $t_1 + \kappa(t_1)$. It is easy to see that in this case T_1 is a close precursor of T_2 in accordance with (4).

4.2 Joint Intensities

In a local sense, we might say that event ∘ *triggers* event × if it increases (for a time) the conditional intensity of ∘, given recent history. For example, let $H(s-)$ again denote the history of the joint process before time s. Then ∘ triggers × if for some $\kappa(s)$, we have

$$\lambda_2(u \mid H(s-) \text{ and } \circ \text{ at } s) = e^{\beta} \lambda_2(u \mid H(s-)) \tag{10}$$

for $s < u < s + \kappa(s)$ $(\beta > 0)$.

It is possible for the two kinds of recurrent events to be triggers for each other.

5. NONPARAMETRIC TESTING FOR ORDER, COMPLETE DATA

The key to a simple non-parametric test for ordering with complete data seems to be to examine the empirical distribution of $T_2 - T_1$. Recall that T_1 is a weak precursor of T_2 if $P(T_2 - T_1 > 0)$ is greater than what it would be under independence of T_1, T_2; T_1 is a close precursor of T_2 if the density function of $T_2 - T_1$, given $T_1 = t_1$ and $T_2 > T_1$, is higher for small positive arguments than the density of $T_2 - t_1$, given $T_2 > t_1$. We illustrate the use of $T_2 - T_1$ with a somewhat artificial example, inspired by the smoking cessation and pregnancy context.

Suppose T_1 (corresponding to becoming pregnant) has a constant hazard rate of 0.1. Suppose that T_2 (corresponding to smoking cessation) has hazard rate 0.05 before T_1, and 0.15 after T_1 for 0.5 time units, reverting to 0.05 thereafter. Take the time origin 0 to be the time of the n th cycle, and simulate the histories of subjects who at the origin have had neither T_1 nor T_2 come to pass (non-pregnant smokers). Suppose we are interested in modelling events occurring between the n-th and $(n+1)$-th cycle.

In one simulation of 10,000 subjects, 145 had $0 < T_1, T_2 < 2$.

To check graphically for T_1 as a (close) precursor of T_2, we could look at distribution (or density) of $T_2 - T_1$, conditional on $0 < T_1, T_2 < 2$. We could compare the empirical or estimated density of $T_2 - T_1$, conditional on $0 < T_1, T_2 < 2$, to the implied conditional density of $T_2 - T_1$, based on the estimated marginals of T_1, T_2, and assuming independence. For the simulated data set, the comparison is shown in Figure 1. The estimated density of $T_2 - T_1$ shows a concentration of mass just above 0.

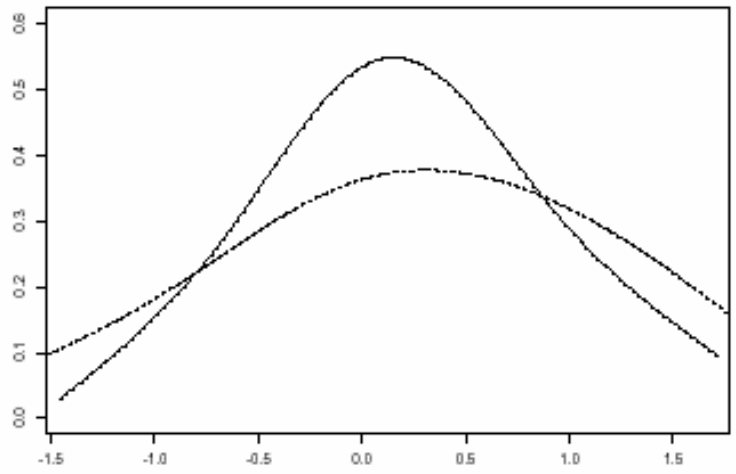

Figure 1: Estimated density of $T_2 - T_1$: empirical density (solid line); density based on independence assumption (dotted line).

A formal nonparametric test for association, which would have power against an alternative where T_1 is a close precursor of T_2, could proceed as follows. Estimate $F_2(.)$ and $F_1(.)$ from the data, as well as the joint distribution of $(T_1, T_2 - T_1)$. The joint distribution of $(T_1, T_2 - T_1)$ need only be estimated to the point of estimating

$$\frac{F_2(t_1 + k(t_1) | T_1 = t_1)}{F_2(t_1 | T_1 = t_1)} = R(t_1) \tag{11}$$

for each t_1 and suitable $\kappa(t_1)$. Compute the test statistic

$$\int (\hat{R}(t_1) - \frac{\hat{F}_2(t_1 + \kappa(t_1))}{\hat{F}_2(t_1)}) d\hat{F}_1(t_1), \tag{12}$$

and compare it with twice its estimated standard error, to test the null hypothesis that its mean is 0.

The estimation of $F_2(.)$, $F_1(.)$, and the joint distribution of $(T_1, T_2 - T_1)$ is not difficult in principle, using methods for distribution and density estimation for complex surveys. See for example Bellhouse and Stafford (1999), Lohr and Buskirk (1999). Resampling methods could be used to assess standard error.

Note: The example is artificial because the assumption of constant hazard for smoking cessation in the absence of pregnancy would be an oversimplification. In fact, when we observe $T_2 < 2$, we are observing the time of beginning of a *cessation attempt* which has lasted long enough to be reported at time 2. Strictly speaking, the model should recognize the possibility of cessation followed by relapse before the end of the period; if this is not included explicitly, we should at least allow the hazard for reported cessation to increase over $[0, 2]$.

6. INTERVAL CENSORING

In longitudinal surveys conducted at widely spaced time points, it is possible for the endpoint of T_1 or T_2, or the time of occurrence of $\circ$ or $\times$, to be *interval censored*, or observed to be within an interval only. This is the case with the time of cessation of smoking in the NPHS. The interviewer asks:

Have you ever smoked cigarettes at all? ...At what age did you stop smoking cigarettes daily? ... Compared to our interview in MONTH/YEAR, you are reporting that you no longer smoke. why did you quit?

If [0, 2] is the period between interviews, and smoking cessation is observed, the time T_2 of smoking cessation is thus observed to be between two endpoints T_{20} and T_{21}, where T_{20} is either 0 or a birthday, and T_{21} is either a birthday or 2. If the subjects are females aged 15 - 45 who are regular smokers and not pregnant at time 0, we can let T_1 be the time they become pregnant, which ideally would be "observed" if it falls before 2.

If we have a semiparametric trigger model along the lines of (9), we can deal with this situation simply by calculating the appropriate likelihood function and using it to estimate β.

For example, suppose $\lambda_1(t)$ and $\lambda_{02}(\tau)$ are respectively the hazard functions for T_1 and for T_2 when T_1 is ∞. Suppose that $\lambda_{12}(u \mid t; \beta)$ is the hazard function for T_2 at u, given $T_1 = t$, so that $\lambda_{12}(u \mid t; \beta)$ could be $e^{\beta}\lambda_{02}(u)$ for u between t and $t + 0.5$, and $\lambda_{02}(u)$ for other u.

Replacing 2 by a general endpoint a_1, and taking t_{20i} to be a_1 if T_2 is unobserved for subject i, we can form the likelihood function $\prod_i L(t_{20i}, t_{21i}, t_{1i})$

where $L(t_{20}, t_{21}, t_1) =$:

1. $A(a_1)$ if $a_1 \le t_1, t_{20}$
2. $A(t_1)\lambda_1(t_1)dt_1 \mathrm{B}_{12}(a_1 \mid t_1; \beta)$ if $0 < t_1 < a_1 = t_{20}$
3. $\int_{t_{20}}^{t_{21}} A(\tau)\lambda_{02}(\tau)exp\{-\int_{\tau}^{a_1}\lambda_1(u)du\}d\tau$ if $0 \le t_{20} < t_{21} \le a_1 < t_1$
4. $\int_{t_{20}}^{t_1} A(\tau)\lambda_{02}(\tau)\exp\{-\int_{\tau}^{t_1}\lambda_1(u)du\}\lambda_1(t_1)d\tau dt_1 + A(t_1)\lambda_1(t_1)dt_1[1 - \mathrm{B}_{12}(t_{21} \mid t_1; \beta)]$ if $0 \le t_{20} < t_1 < t_{21} \le a_1$
5. $A(t_1)\lambda_1(t_1)dt_1[\mathrm{B}_{12}(t_{20} \mid t_1; \beta) - \mathrm{B}_{12}(t_{21} \mid t_1; \beta)]$ if $0 < t_1 < t_{20} < t_{21} \le a_1$
6. $\int_{t_{20}}^{t_{21}} A(\tau)\lambda_{02}(\tau)exp\{-\int_{\tau}^{t_1}\lambda_1(u)du\}\lambda_1(t_1)d\tau dt_1$ if $0 < t_{20} < t_{21} \le t_1 < a_1$

and

$$A(t) = exp\{-\int_0^t [\lambda_1(u) + \lambda_{02}(u)]du\},$$

$$B_{12}(\tau \mid t; \beta) = \exp\{-\int_t^{\tau}\lambda_{12}(u \mid t; \beta)du\}.$$

More realistic models would allow the hazard λ_1 to depend on variates such as marital status and age, and the hazards λ_{02} and λ_{12} to depend on variates such as education level and some measure of strength of addiction.

For nonparametric testing in the spirit of (12), we need to adapt the estimation of the survivor functions and the joint distribution of T_1, $T_2 - T_1$ given $0 < T_1, T_2 < a_1$ to the case where T_2 is interval censored. A

straightforward approach would construct a bivariate density estimate for T_1, T_2 on the rectangle $(0 < T_1, T_2 < a_1)$, along the lines of the method proposed by Duchesne and Stafford (2002) and using kernels oriented parallel to the line $t_1 = t_2$; then integrate to produce the numerator and denominator of $\hat{R}(t_1)$ for each t_1, as well as the other quantities in (12).

ACKNOWLEDGEMENTS

This research is supported by a Research Grant from NSERC. Discussions with J. F. Lawless are gratefully acknowledged.

REFERENCES

Bellhouse, D. and Stafford, J. (1999), "Density Estimation from Complex Surveys", Statistica Sinica, 9, pp. 407-424.

Bradford Hill, A. (1937), Principles of Medical Statistics, Arnold.

Bradford Hill, A. (1965), "The Environment and Disease: Association or Causation", Proceedings of the Royal Society of Medicine, 58, pp. 295-300.

Cox, D. R. (1992), "Causality: Some Statistical Aspects", Journal of the Royal Statistical Society, Series A, 155, pp. 291-301.

Duchesne, T. and Stafford, J. E. (2002), "A Kernel Density Estimate for Interval Censored Data",

Lohr, S. and Buskirk, T. (1999), "Density Estimation with Complex Survey Data", Proceedings of the Survey Methods Section, Statistical Society of Canada, pp. 27-32.