Event History Analysis

REPEATED EVENTS

Contributors: Paul D. Allison
Book Title: Event History Analysis
Chapter Title: "REPEATED EVENTS"
Pub. Date: 1984
Access Date: March 17, 2015
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9780803920552
Online ISBN: 9781412984195
DOI: http://dx.doi.org/10.4135/9781412984195.n6
Print pages: 51-58

http://dx.doi.org/10.4135/9781412984195.n6
[p. 51 ↓]

REPEATED EVENTS

Most events studied by social scientists are repeatable, and most event history data contain repeated events for each individual. Examples include job changes, births, marriages, divorces, arrests, convictions, and visits to a physician. Unfortunately, the enormous literature on event history methods that has come out of biostatistics contains only a handful of articles on the analysis of repeated events (e.g., Gail, Santner, and Brown, 1980; Prentice, Williams, and Peterson, 1981). As already noted, this is a consequence of the fact that the events of greatest interest in biomedical research are deaths. While sociologists (Tuma, Hannan, and Groeneveld, 1979) and economists (Flinn and Heckman, 1982a, 1982b) have made some progress in this area, there is still much to be done in the way of developing methods that are suitable for repeated events.

One approach that is sometimes appropriate is to conduct a separate analysis for each successive event using any of the methods already discussed. In a study of marital fertility, for example, one could estimate a model for the interval between marriage and first birth, a second model for the interval between first and second birth, and so on. This approach requires no special assumptions, and is especially useful if one expects the model to differ from one event to another. On the other hand, if the process is essentially the same across successive events, doing a separate analysis for each event is both tedious and statistically inefficient.

A Simple Approach

In this chapter we shall focus on a second approach that avoids these difficulties by treating each interval between events for each individual as a separate observation. These intervals (sometimes referred to as spells) are then pooled over all individuals. At this point any of the methods described in the previous chapters can be applied. Although this approach is not entirely satisfactory from the viewpoint of statistical theory,

we shall postpone a discussion of those problems until later in this chapter. There are also a number of possible complications not discussed in earlier chapters.

To simplify the discussion, let us assume that the repeated events are of a single kind. (In the next chapter we shall consider models that incorporate both multiple kinds of events and repeated events.) We begin by extending the empirical example discussed in Chapter 5. Recall **[p. 52 ↓]** that the sample consisted of approximately 1000 inmates released from Georgia state prisons who were observed for one year after their release. In the previous analysis, the event of interest was the first arrest that occurred after release. As Table 5 shows, however, many of the subjects were arrested more than once during the one-year follow-up period, and to ignore the later arrests seems wasteful of information. It also raises the question of whether the causal process differed for earlier and later arrests.

TABLE 5 Frequency Distribution for Arrests

TABLE 5
Frequency Distribution for Arrests

Number of Arrests	*Number of Persons*
0	622
1	213
2	85
3	25
4	9
5	5
6	2

To incorporate these additional arrests into an event history analysis, let us divide each individual's one-year follow-up period into intervals, using the observed arrests as the dividing points. Consider, for example, a person with two arrests that occurred at times marked by Xs on the line below:

With two arrests there are three intervals, the last of which is censored by the end of the observation period. Similarly, a person with four arrests would have five intervals, the last of which is censored. Thus, every individual has exactly one censored interval, and may also have one or more uncensored intervals. For this sample, the 961 persons had a total of 1492 intervals, and, of course, 961 of those were censored.

Treating each of those intervals as a separate observation and pooling all intervals, we now reestimate the proportional hazards model corresponding to column 1 of Table 4. Results are shown in Table 6, column 1. Notice that the effective number of observations is increased by about 50 percent, and the number of observed arrests is increased by over 60 percent.**[p. 53 ↓]** It is reasonable to expect, then, that the new estimates should have smaller standard errors and, hence, larger t-statistics. Indeed, although the basic pattern of results is the same, we do find somewhat larger t-statistics when all the arrests are included. In fact, the positive effect of financial aid is now marginally significant by a one-tailed test.

TABLE 6 Estimates of Proportional Hazards Models for Repeated Arrests

TABLE 6
Estimates of Proportional Hazards Models for Repeated Arrests

	All Arrests				*Second or Later Arrests*	
	1		*2*		*3*	
Explanatory Variables	*b*	*t*	*b*	*t*	*b*	*t*
Education	−.008	−.35	−.010	.45	.020	.52
Financial aid (D)[a]	.150	1.69	.136	1.54	.142	.95
Imprisoned for crime against person (D)	.173	1.25	.153	1.11	.207	.92
Imprisoned for crime against property (D)	.367	3.17**	.336	2.88**	.090	.45
Number of convictions for crimes against persons	.051	.10	−.002	.00	.162	.75
Number of convictions for crimes against property	.183	3.27**	.155	2.73**	.061	.66
Paroled (D)	.341	3.65**	.312	3.31**	.297	1.75
Male (D)	.054	.26	.076	.37	−.134	.45
Age at earliest arrest	−.043	4.20**	−.038	3.73**	−.022	1.27
Married (D)	.160	1.50	.151	1.42	.293	1.71
Age at release	−.008	1.17	−.008	1.14	−.003	.30
Number of prior intervals	–	–	.197	2.59**	.087	.96
Time since release	–	–	.017	2.05*	−.001	1.01
N or arrests	549		549		203	
N	1492		1492		531	

a. (D) indicates dummy variable.
*Significant at .05 level.
**Significant at .01 level.

Problems with Repeated Events

While the method just described is straightforward and intuitively appealing, it requires that one make a number of assumptions that may well be problematic. First, one must assume that the dependence of the hazard on time since last event has the same form for each successive event. Recall that in a proportional hazards model,

$$\log h(t) = a(t) + b_1x_1 + b_2x_2 + \ldots \qquad [18]$$

[p. 54 ↓] where t is the length of time since the last event and a(t) is an unspecified function of time. Even though it is unspecified, a(t) must be the same function for the first arrest, the second arrest, and so on. Or if one assumes that intervals have a Weibull distribution, that distribution must have the same shape parameter for each successive interval. If there is reason to be suspicious of this assumption, the proportional hazards model can be modified to make it unnecessary. The basic idea is to let the function a(t) be different for each successive interval, while forcing the b coefficients to be the same. Such a model can be readily estimated using the method of stratification discussed in Chapter 4.

A second assumption implicit in this method is that, for each individual, the multiple intervals must be statistically independent. In general, we would expect that people who are frequently arrested (i.e., have short intervals) will continue to be frequently arrested. This does not violate the assumption of independence, so long as that dependence is fully accounted for by the explanatory variables included in the model.

In most cases, however, there will be good reason to think that the independence assumption is false, at least to some degree. The consequences of violating this assumption have not been studied, but analogies with linear regression suggest that (a) the coefficient estimates will still be asymptotically unbiased and (b) standard error estimates will be biased downward. Work is now being done on ways to relax the independence assumption by introducing a random disturbance term that is correlated across intervals (Flinn and Heckman, 1982a, 1982b). This approach has not progressed to the point where the new methods can be generally recommended, however.

In the meantime, there are some things that can be done to minimize the consequences of violating the independence assumption. The basic idea is to include in the model additional explanatory variables that tap characteristics of the individual's prior event history. The simplest such variables are the number of events prior to the interval in question, and the length of the previous interval, set to zero when no previous interval is observed.

Another approach to the problem of dependence is to modify the estimated standard errors so that they reflect the number of individuals rather than the number of intervals. Let n and N be the number of individuals and the number of intervals, respectively.

The standard errors may be adjusted upward by multiplying each one by the square root of N/n. Similarly, the t-statistics may be adjusted downward by multiplying each one by the square root of n/N. The rationale for this adjustment is that, if intervals are highly dependent, the multiple intervals**[p. 55 ↓]** for a single individual are redundant —an individual with many intervals is not contributing much more information than an individual with just one interval. The approach is highly conservative, however, and probably results in overestimates of the true standard errors.

A third limitation of models for repeated events considered thus far is that the hazard rate is expressed as a function of the time since the last event. While this is by far the most common specification, there are often situations in which it is more plausible to let the hazard vary as a function of age or time since some common starting point. In studies of fertility, for example, age may have stronger effects on the hazard for a birth than length of time since the previous birth. We have discussed the problem of starting points in Chapter 4, but some additional remarks are in order here. First, the question of origin of the time scale is *always* ambiguous in the case of repeated events and should always be given careful consideration. Second, models for repeated events in which the hazard depends on time since some fixed starting point may be inconvenient to estimate. Consider, for example, the proportional hazards model in equation 18 where we now consider t to be time since release from prison rather than time since the last event. Such a model can, in principle, be estimated by the partial likelihood method, but the commonly available partial likelihood programs do not have that capability.[7] See Prentice, Williams, and Peterson (1981) for further details. Perhaps the best approach, at present, is to include age or time since some other point as an explanatory variable in models that allow the hazard to vary with time since the last event.

Extending the Recidivism Example

Let us now return to the recidivism example to incorporate some of the new possibilities just discussed. In panel 2 of Table 6, estimates are given for a model that includes number of prior arrests and length of time from release to the beginning of each interval. Both of these new variables are statistically significant. The positive effect of number of prior arrests indicates, as expected, that those with many arrests have a higher

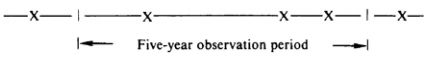
hazard for arrest at subsequent points in time. The positive effect of time since release indicates a tendency for the hazard to increase over the one-year observation period. The inclusion of these variables somewhat attenuates the coefficients and t-statistics for the other variables in the model. In fact, the t-statistics are about the same as those for the model which only examined the first arrest. Some attenuation is to be expected since the violation of the independence assumption is likely to lead to inflated t-statistics.

[p. 56 ↓] When corrections are introduced for possible violations of assumptions, it appears that not much has been gained by analyzing all arrests rather than just the first arrest. This is surprising since the inclusion of the additional arrests ought to have yielded diminished standard errors and, hence, increased t-statistics. One possible explanation is that the causal process may be somewhat different for arrests after the first. To examine this possibility, the second model was reestimated after excluding all the intervals from release to first arrest. This left 531 intervals of which 231 ended in arrest. Results are shown in panel 3 of Table 6. With the exception of parole status, none of the explanatory variables even approaches statistical significance. It is reasonable to expect *some* decline in significance level since the effective sample size has been reduced greatly. Nevertheless, the coefficients of the formerly significant variables are also greatly attenuated, suggesting that there has been a real decline in the effects of these variables for later arrests. We shall not speculate on the reasons for this decline.

Left Censoring

Before leaving the topic of repeated events, let us consider one further problem that is quite common but did not occur in this particular example. The problem is often referred to as “left censoring,” but it is worth noting that biostatisticians mean something quite different when they use this term.[8]

Suppose that a sample of people is observed over a five-year period, and the event of interest is a job change. During that five-year period, the pattern for a particular individual with three job changes might look like this:



where X denotes the occurrence of a job change and the two extreme Xs (to the right and left of the vertical lines) refer to events that are not observed. The three observed events divide the five-year period into four intervals. The last interval is clearly censored and the middle two intervals are clearly uncensored. The first interval is problematic, however. Although it ends with an event, the interval is still censored because the time of the preceding event is unknown.

[p. 57 ↓] The consequences of left censoring depend on the model being estimated. If the model specifies a hazard rate that does not depend on time (an exponential model in the continuous case), there is no problem whatever. One simply treats the initial censored interval as if it began at the beginning of the observation period. Similarly, there is no inherent difficulty if the hazard rate depends on age, except for the computational problem noted above.

Dependence on the time since the last event poses serious problems, however, because the time since the last event is not known. Treating the interval as if it began at the beginning of the observation period will undoubtedly introduce some bias. Flinn and Heckman (1982a, 1982b) discuss several possible solutions to this problem, but all are computationally cumbersome and depend on somewhat arbitrary assumptions. The safest approach is simply to discard the initially censored intervals. While this represents a loss of information, it should not lead to any biases.

http://dx.doi.org/10.4135/9781412984195.n6