Event History Analysis

PROPORTIONAL HAZARDS AND PARTIAL LIKELIHOOD

Contributors: Paul D. Allison
Book Title: Event History Analysis
Chapter Title: "PROPORTIONAL HAZARDS AND PARTIAL LIKELIHOOD"
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.n4
Print pages: 34-43

http://dx.doi.org/10.4135/9781412984195.n4

PROPORTIONAL HAZARDS AND PARTIAL LIKELIHOOD

The methods discussed and applied in Chapter 3 represent a tremendous advance over ad hoc approaches to event history data, but they still have some disadvantages. First, it is necessary to decide how the hazard rate depends on time, and there may be little information on which to base such a choice. Moreover, if the hazard function is believed to be nonmonotonic, it may be difficult to find a model with the appropriate shape. Much experience with these models suggests that the coefficient estimates are not terribly sensitive to the choice of the hazard function, but one can never be sure what will happen in any particular situation. The second, and perhaps the more serious, problem is that these models do not allow for explanatory variables whose values change over time. While it is possible to develop fully parametric models that include time-varying explanatory variables (Tuma, 1979; Flinn and Heckman, 1982a, 1982b) estimation of these models is somewhat cumbersome.

Both these problems were solved in 1972 when David Cox, a British statistician, published a paper entitled “Regression Analysis and Life**[p. 34 ↓]** Tables,” in which he proposed a model and an estimation method that have since become extremely popular, especially in biomedical research.

The Proportional Hazards Model

Commonly referred to as the “proportional hazards model,” Cox's model is a simple generalization of the parametric models we have just considered. We shall postpone, for the moment, a consideration of models with time-varying explanatory variables. For two time-constant variables, the model may be written as

$$\log h(t) = a(t) + b_1 x_1 + b_2 x_2 \qquad [10]$$

where a(t) can be any function of time. Because this function does not have to be specified, the model is often described as partially parametric or semiparametric. It is called the proportional hazards model because for any two individuals at any point in time, the ratio of their hazards is a constant. Formally, for any time t, h
i

(t)/h
j

(t) = c where i and j refer to distinct individuals and c may depend on explanatory variables but not on time. Despite the name, this is not a crucial feature of Cox's model because the hazards cease to be proportional as soon as one introduces time-varying explanatory variables.

Partial Likelihood

Again, it is easy to write down such models but difficult to devise ways to estimate them. Cox's most important contribution was to propose a method called partial likelihood which bears many similarities to ordinary maximum likelihood estimation. Mathematical details on partial likelihood are given in Appendix A, but some general properties can be mentioned here. The method relies on the fact that the likelihood function for data arising from the proportional hazards model can be factored into two parts: One factor contains information only about the coefficients b
1

and b
2

; the other factor contains information about b
1

, b
2

, and the function a(t). Partial likelihood simply discards the second factor and treats the first factor as though it were an ordinary likelihood function. This first factor depends only on the *order* in which events occur, not on the exact times of occurrence.[5]

The resulting estimators are asymptotically unbiased and normally distributed. They are not fully efficient because some information is lost by ignoring the exact times of event occurrence. But the loss of efficiency is usually so small that it is not worth worrying about (Efron, 1977).

[p. 35 ↓] It is difficult to exaggerate the impact of Cox's work on the practical analysis of event history data. In recent years, his 1972 paper has been cited well over 100 times a year in the world scientific literature. In the judgment of many, it is unequivocally the best all-around method for estimating regression models with continuous-time data.

Other methods may be more appropriate in cases where a major substantive concern is with the dependence of the hazard on time. For example, the principle of cumulative inertia suggests that the longer an individual is in a particular state, the less likely he is to leave that state (McGinnis, 1968). Such a hypothesis could not be tested under partial likelihood estimation. In most cases, however, the main concern is with the effects of the explanatory variables, and the dependence on time is of little interest. As noted in Chapter 3, moreover, testing hypotheses about the effect of time on the hazard is difficult under any model because unmeasured sources of heterogeneity usually contaminate that effect.

Computer programs for partial likelihood estimation for the proportional hazards model are now widely available as part of the SAS (SAS Institute, 1983) and BMDP (Dixon, 1981) statistical packages. Other publicly available programs to do partial likelihood are RATE (Tuma, 1979) and SURVREG (Preston and Clarkson, 1983). The SAS supplementary procedure PHGLM is very easy to use but does not allow for time-varying explanatory variables. Neither does the proportional hazards model in RATE.

Partial Likelihood Applied to an Empirical Example

Let us return to the criminal recidivism data of Rossi et al. (1980) to see what happens when a proportional hazards model is estimated. Using the same explanatory variables as for the exponential regression model, estimates were obtained using the SAS procedure PHGLM. (A program listing is given in Appendix B.) Results are reported in panel 2 of Table 3. Both the coefficient estimates and the ratios of the estimates to their standard errors (t-statistics) are almost identical to those produced by the exponential regression model. This should not be too surprising since the exponential model is just a proportional hazards model in which the arbitrary function a(t) is fixed at a constant value. The fact that the estimates are so similar suggests that the hazard for an arrest does not change much over the 12-month period. Given these results, there would be no point in estimating a Weibull regression model; since the Weibull model falls between the exponential and**[p. 36 ↓]** proportional hazards models in generality, the estimates would hardly vary from those in panels 1 and 2 of Table 3.

Time-Varying Explanatory Variables

The proportional hazards model can be extended easily to allow for explanatory variables that change in value over time. A model with two explanatory variables, one constant and one varying over time, may be written as

$$\log h(t) = a(t) + b_1 x_1 + b_1 x_2(t) \qquad [11]$$

where, as before, a(t) may be any function of time. This model says that the hazard at time t depends on the value of x
2

at the same time t. In some cases, however, there may be reason to believe that there is a lag between a change in a variable and the effect of that change on the hazard. For example, if one is interested in the effect of a bout of unemployment on the hazard of

divorce, it might be plausible to suspect that there is a lag between the loss of a job and an increase in the hazard. If the suspected lag is two months (and time is measured in months) the model can be modified to read

$$\log h(t) = a(t) + b_1x_1 + b_2x_2(t - 2) \qquad [12]$$

With or without lags, models with time-varying explanatory variables can be estimated using the partial likelihood method described earlier. The derivation of the partial likelihood function is essentially the same with time-varying explanatory variables, but the computer algorithms for constructing and maximizing that likelihood function are more complex. Hence, not all programs for partial likelihood estimation will handle time-varying explanatory variables (sometimes referred to as “time-dependent covariates”).

Returning to the recidivism example, we noted earlier that the variable “weeks employed during the first three months after release” is merely a substitute for what is actually a time-varying explanatory variable. What one would ideally like to know is how the hazard is affected by employment status at any given point in time during the one-year follow-up. This question can be answered because the data set includes information on whether the individual was employed during each of the 52 weeks of observation. Using BMDP2L, a proportional hazards model was estimated that included a dummy variable for**[p. 37 ↓]** employment status as a time-varying explanatory variable. (See Appendix B for program listing.) Results are shown in panel 3 of Table 3.

For the most part the results are quite similar to those in panels 1 and 2. The big difference is in the effect of employment status, which is now clearly the most important variable in the model. Exponentiating the coefficient of -1.397 yields .25, which says that the hazard of arrest for those who were working was only one fourth the hazard of those who were not working.

Problems with Time-Varying Explanatory Variables

A word of warning is in order here. Regardless of the computer program, estimation of proportional hazards models with time-varying explanatory variables can enormously increase computational costs. In this example, for instance, the CPU time increased by a factor of 10 with the inclusion of just one time-varying explanatory variable. Moreover, setting up the model may not be straightforward. With the exception of variables that are very simple functions of time itself, BMDP2L requires the inclusion of a FORTRAN subroutine to define the time-varying variables. Procedures for doing this are not well documented.

Another possible complication in the estimation of models with time-varying explanatory variables involves the frequency with which those variables are measured. Strictly speaking, estimation of such models requires that for each time that an event occurs, values of the explanatory variables must be known for all individuals at risk at that time. Thus, if an event occurred at time 10, and 15 individuals were at risk at that time, the values of the explanatory variables at time 10 must be known for all 15 individuals. Typically, that would require that the explanatory variables be measured continuously over time.

In practice, however, time-varying explanatory variables are usually measured at regular intervals. In the example just considered, employment status was known for each week of observation. That created no problem because the time of arrest was measured in weeks. Difficulties arise when time of event occurrence is measured more precisely than the interval at which the explanatory variables are measured. For example, event times may be measured in days but the values of the explanatory variables may be known only at the beginning of each month.

In such cases, some ad hoc procedure will be necessary to estimate the values of the explanatory variables at the times of events. The simplest approach is to use the value closest in time to the event time as the**[p. 38 ↓]** estimated value. A better method is to use linear interpolation, which is equivalent to weighted averaging. Suppose, for

example, that the values of an explanatory variable x are known at time 10 and time 20 but an event occurs at time 13. An estimate of x(13) is given by x(10)(.7) + x(20)(.3). For a discussion of these and other methods, see Tuma (1982).

Adequacy of the Proportional Hazards Model

Many researchers worry about whether their data satisfy the proportional hazards assumption. For those with such concerns, there are ways of both assessing the validity of this assumption and altering the model to correct for violations. Before discussing these methods, however, let us consider the possibility that these worries may be exaggerated. As models go, the proportional hazards model is extraordinarily general and nonrestrictive—the main reason for its great popularity. Even when the proportional hazards assumption is violated, it is often a satisfactory approximation. Those who are concerned about misspecification would usually do better to focus on the possibilities of omitted explanatory variables, measurement error in the explanatory variables, and nonin-dependence of censoring and the occurence of events.

With that in mind, let us proceed to talk about nonproportional hazards. What does it mean to say that hazards are not proportional? As noted earlier, models with time-varying explanatory variables do not have proportional hazards, but there are other ways in which this can occur. When the dependent variable is the logarithm of the hazard, the hazards are not proportional if there is an *interaction* between time and one or more explanatory variables, e.g.,

$$\log h(t) = a(t) + bx + cxt \qquad [13]$$

This model differs from the usual model by including the product of x and t as one of the explanatory variables. If c is positive, we can say that the effect of time on the hazard increases linearly as x increases. Alternatively, we can say that the effect of x on the hazard goes up linearly with time. Hence, the hazards are not proportional if the effect of some explanatory variable on the hazard is different at different points in time.

There are other ways of expressing interaction, of course, but the partial likelihood method can estimate a broad class of these by including additional variables in the model. Specifically, it can estimate models of the form

$$\log h(t) = a(t) + bx + c\,g(x,t) \qquad [14]$$

[p. 39 ↓] where g(x, t) is some nonlinear function of x and t with known parameters. One simply computes g(x, t) for each event time t and includes this in the model as a time-varying explanatory variable. Thus, one very good way of testing for nonproportionality is to estimate models of this form, and then test whether the coefficient c differs significantly from zero. If it does, then one has already solved the problem by estimating the extended model.

A limitation of this approach is that not all partial likelihood programs allow time-dependent explanatory variables. Moreover, estimation of such models tends to be expensive. There is a much cheaper graphical method that works well when time interacts with a categorical (nominal) variable, or one that can be treated as categorical. In this method, a certain function of time (the log-log survival function) is plotted for subsamples corresponding to the different levels of the categorical variable. If the categorical variable is sex, for example, one plot is produced for males and another plot is produced for females. These plots should be roughly parallel if the proportional hazards assumption is satisfied. (For further information see Lawless, 1982, and the BMDP2L manual.)

If this graphical test provides evidence against the proportional hazards assumption, there is a method called “stratification” that allows for nonproportionality. It is much cheaper than using time-varying explanatory variables. The basic idea is to divide the sample into strata according to the values of the categorical variable that interacts with time. A separate model is then postulated for each stratum.

Suppose, for example, that males and females are thought to have nonproportional hazards. We then specify two models:

$$\text{males: } \log h(t) = a_1(t) + b_1x_1 + b_2x_2 + \ldots$$
$$\text{females: } \log h(t) = a_2(t) + b_1x_1 + b_2x_2 + \ldots \qquad [15]$$

These models share the same set of b coefficients but each has a different (but unspecified) function of time. Both models can be estimated simultaneously using the partial likelihood method. Stratification is available in the BMDP2L program, the 1983 version of the SAS PHGLM procedure, and in the SURVREG program (Preston and Clarkson, 1983).

Another approach to checking the adequacy of the model is the examination of residuals. Methods for calculating residuals from the proportional hazards model have been proposed by Kay (1977) and Schoenfeld (1982). Kay's formula has been incorporated into the BMDP2L program, which also plots those residuals in such a way that deviations from a straight line represent a failure of the model. Unfortunately,**[p. 40 ↓]** a recent investigation (Crowley and Storer, 1983) raises questions about the diagnostic value of such plots.

Those accustomed to multiple regression will undoubtedly wish that there were something like R^2 for proportional hazards models. Harrel (1980) has developed an R^2 analog for these models, and has incorporated it into the SAS supplementary procedure PHGLM. From his description, however, it is not clear that this is the most appropriate analog, and others may yet be developed. In any case, it must be emphasized that such a statistic does *not* measure how well the assumptions of the model are satisfied by the data. As in ordinary multiple regression, a low R^2 is quite compatible with a model that is perfectly specified, and a high R^2 can be found for models that are grossly misspecified. Such statistics only tell one how much variation in the dependent variable is attributable to variations in the explanatory variables.

Choice of Origin of the Time Scale

One aspect of model choice that has been implicit to this point is the question of when time begins. While this question is relevant to both parametric and semiparametric

models, a discussion has been postponed until now because the proportional hazards model offers a wider range of possibilities. In the recidivism example, the origin of the time scale was relatively unambiguous. The date of release is a natural starting point for calculating time of first arrest after release. Similarly, if one is estimating models for the hazard of divorce, the date of marriage is a natural starting point for calculating time of divorce.

There are many cases, however, in which the origin of the time scale will not be so clear. Even in seemingly straightforward examples, moreover, there is room for disagreement. In formulating a proportional hazards model for recidivism, for instance, one could let the hazard be a function of the person's age or calendar time rather than time since release. Age and calendar time are also possible time scales for the hazard for divorce.

There would be no difficulty in estimating such models by partial likelihood, but would it be reasonable to do so? That depends on substantive considerations. If the hazard is known to depend strongly on age but only weakly on time since some other starting point, then age would probably be the most appropriate way to define the time scale. Or if the hazard is thought to vary greatly with historical conditions that affect all sample members in the same way, then calendar time might be the best time scale.

[p. 41 ↓] In theory, one could formulate and estimate proportional hazards models in which the hazard depended arbitrarily on two or more time scales. In practice, this requires very large samples or special conditions described by Tuma (1982). Even when this approach is not possible, however, one can always explicitly introduce different time scales as explanatory variables. For example, in estimating a proportional hazards model for divorce in which the hazard varies arbitrarily with duration of marriage, one could also include calendar year, age of husband, and age of wife as explanatory variables. If any of these variables has a nonlinear effect (on the log of the hazard), it is necessary to specify the variable as a time-varying explanatory variable. When the effect is linear (on the log of the hazard), however, it is sufficient to measure the variable at the beginning of the principal time scale, in this case the beginning of the marriage.[6] Another example is provided by Table 3 where age at release is included as an explanatory variable.

Partial Likelihood for Discrete-Time Data

The discussion of the partial likelihood method has assumed that time is measured on a continuous scale and that, as a consequence, two events cannot occur at exactly the same time. In practice, time is always measured in discrete units, however small, and many data sets will contain “ties”—two or more individuals experiencing events at apparently the same time. Thus, in the recidivism example where time was measured in weeks, there were several weeks in which two or more persons were arrested.

To handle such data, both the model and the partial likelihood method must be modified to some degree. The model proposed by Cox (1972) for data with ties is just the logit-linear model of equation 3 in Chapter 1. This model is attractive because, as the discrete-time units get smaller and smaller, it converges to the proportional hazards model (Thompson, 1977).

The method of partial likelihood may be used to estimate this model but, if the number of ties is at all large, the computational requirements are gargantuan. To avoid this, a number of approximations have been proposed, the most widely accepted of which is Breslow's (1974). This formula has been incorporated into most programs for partial likelihood estimation including all those mentioned here. When there are no tied data, Breslow's formula reduces to the usual partial likelihood for continuous-time data. Thus, such programs can be used for either continuous- or discrete-time data, and there is usually no need for the researcher to be concerned about the occurrence of ties.

[p. 42 ↓] This approach is not fool-proof, however. If the number of events occurring at most time points is large relative to the number at risk (say, 50 percent or more), the Breslow approximation will be poor (Farewell and Prentice, 1980). In such situations, it would be better to use the discrete-time estimation method described in Chapter 1.

http://dx.doi.org/10.4135/9781412984195.n4