Event History Analysis

A DISCRETE-TIME METHOD

Contributors: Paul D. Allison
Book Title: Event History Analysis
Chapter Title: "A DISCRETE-TIME METHOD"
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.n2
Print pages: 15-23

http://dx.doi.org/10.4135/9781412984195.n2

A DISCRETE-TIME METHOD

This chapter introduces discrete-time methods for unrepeated events of a single kind. While this is among the simplest situations, it involves**[p. 15 ↓]** many of the fundamental ideas that are central to more complex forms of data. At the same time, the method to be described is eminently practical and can be applied in a great many situations. It can also be generalized to allow for repeated events of different kinds (Allison, 1982).

TABLE 1 Distribution of Year of Employer Change, 200 Biochemists

TABLE 1
Distribution of Year of Employer Change, 200 Biochemists

Year	*Number Changing Employers*	*Number at Risk*	*Estimated Hazard Rate*
1	11	200	.055
2	25	189	.132
3	10	164	.061
4	13	154	.084
5	12	141	.085
>5	129		
Total	200	848	

A Discrete-Time Example

Let us begin with an empirical example. The sample consists of 200 male biochemists who received their doctorates in the late 1950s and early 1960s, and who at some point in their careers were assistant professors in graduate university departments. For a detailed description of the sample, see Long, Allison, and McGinnis (1979). They were observed for a maximum of five years, beginning with the first year of their first positions as assistant professors. The event of interest is the first change of employers to occur after entry into the initial position. Thus, even though we are dealing with what is, in principle, a repeatable event, we define it to be nonrepeatable by restricting our attention to the first employer change. This is an appropriate strategy if one suspects that the process of leaving the first job differs from that of later jobs.

These events are recorded in discrete time since we know only the year in which the employer change occurred, not the exact month and day. In theory it would be desirable

to distinguish voluntary and involuntary changes, but that information is unavailable. Hence, we are dealing with events of a single kind. Table 1 shows the number of biochemists who changed employers in each of the five years. Of the 200 cases, 129 did not change employers during the observation period and are therefore considered to be censored.

[p. 16 ↓] Our goal is to estimate a “regression” model in which the probability of an employer change in a one-year period depends on five explanatory variables. Two of these variables describe the initial employing institution and are assumed to be constant over time: a measure of the prestige of the employing department (Roose and Andersen, 1970) and a measure of federal funds allocated to the institution for biomedical research. Three variables describing individual biochemists were measured annually: cumulative number of published articles, number of citations made by other scientists to each individual's life work, and academic rank coded 1 for associate professor and 0 for assistant professor. (Although all the biochemists began the observation period as assistant professors, some were later promoted.) Thus, we have both constant and timevarying explanatory variables.

The Discrete-Time Hazard Rate

We now proceed to the development of the model. A central concept in event history analysis is the *risk set*, which is the set of individuals who are at risk of event occurrence at each point in time. For the sample of biochemists, all 200 are at risk of changing employers during the first year and thus the entire sample constitutes the risk set in that year. Only 11 of them actually did change employers in that year, and these 11 are no longer at risk during the second year. (They may be a risk of a second employer change but we are only considering the first such change.) Hence, at the end of each year the risk set is diminished by the number who experienced events in that year. In Table 1, for example, we see that the number in the risk set declines from 200 in year 1 to 141 in year 5.

The second key concept is the *hazard rate*, sometimes referred to as simply the hazard or the rate. In discrete time, the hazard rate is the probability that an event will occur at a particular time to a particular individual, given that the individual is at risk at that time.

In the present example, the hazard is the probability of making a first job change within a particular year for those who have not yet changed jobs. It is important to realize that the hazard is an unobserved variable, yet it controls both the occurrence and the timing of events. As such, it is the fundamental dependent variable in an event history model.

If it is assumed that the hazard rate varies by year but is the same for all individuals in each year, one can easily get estimates of the hazard rate: In each year, divide the number of events by the number of individuals at risk. For example, in the second year, 25 biochemists changed employers out of 189 who were in the risk set. The estimated hazard is then 25/189 = .132. Estimates for the other years are shown in **[p. 17 ↓]** the last column of Table 1. There does not appear to be any tendency for the hazard of an employer change to increase or decrease with time on the job. Note also that, because the risk set steadily diminishes, it is possible for the hazard rate to increase even when the number who change employers declines. The estimated hazard rate in year 3, for example, is greater than the hazard rate in year 1 even though more persons changed employers in year 1.

A Logit Regression Model

The next step is to specify how the hazard rate depends on explanatory variables. We shall denote the hazard by P(t), the probability that an individual has an event at time t, given that the individual is still at risk of an event at time t. For simplicity, let us suppose that we have just two explanatory variables: x
1

which is constant over time, and x
2

(t), which has a different value at each time t. For the biochemistry example, x
1

might be prestige of employing department and x
2

(t) might be cumulative number of publications in year t.

As a first approximation, we could write P(t) as a linear function of the explanatory variables:

$$P(t) = a + b_1x_1 + b_2x_2(t) \qquad [1]$$

for t = 1,…, 5. A problem with this specification is that P(t), because it is a probability, cannot be greater than one or less than zero, while the right-hand side of the equation can be any real number. Such a model can yield impossible predictions that create difficulties in both computation and interpretation. This problem can be avoided by taking the logit transformation of P(t):

$$\log(P(t)/(1 - P(t)) = a + b_1x_1 + b_2x_2(t) \qquad [2]$$

As P(t) varies between 0 and 1, the left-hand side of this equation varies between minus and plus infinity. There are other transformations that have this property, but the logit is the most familiar and the most convenient computationally. The coefficients b
1

and b
2

give the change in the logit (log-odds) for each one-unit increase in x
1

and x
2

, respectively.

The model is still somewhat restrictive because it implies that the only changes that occur in the hazard over time are those which result directly from changes in x
2

, the time-varying explanatory variable. In most cases, there are reasons to suspect that the hazard changes autonomously**[p. 18 ↓]** with time. With job changes, for example, one might expect a long-term decline in the hazard simply because individuals become more invested in a job and, hence, the costs of moving increase. On the other hand, untenured academic jobs might show an increase in the hazard after about six years when many individuals are denied promotion.

With the discrete-time model, one can allow for *any* variation in the hazard by letting the intercept *a* be different at each point in discrete time. Thus we can write

$$\log(P(t)/(1 - P(t)) = a(t) + b_1x_1 + b_2x_2(t) \qquad [3]$$

where a(t) refers to five different constants, one for each of the five years that are observed. As we shall see, these constants are estimated by including a set of dummy variables in the specified model.

Estimating the Model

The next problem is to estimate the parameters b
1

, b
2

, and the five values of a(t). As with the models we shall consider, estimation is best done by maximum likelihood or some closely related procedure. The principle of maximum likelihood is to choose as coefficient estimates those values which maximize the probability of observing what has, in fact, been observed. To accomplish this one must first express the probability of the observed data as a function of the unknown coefficients. Then one needs a computational method to maximize this function. Both of these steps are somewhat difficult mathematically, and neither is crucial to a good working knowledge of how to estimate the model. For further details, the interested reader should consult Allison (1982). Happily, estimation reduces to something that is now familiar to many who work with dichotomous dependent variables.

In practice the procedure amounts to this: For each unit of time that each individual is known to be at risk, a separate observational record is created. In our biochemistry example where individuals are persons and time is measured in years, it is natural to refer to these observations as person-years. Thus biochemists who changed employers in year 1 contribute 1 person-year each. Those who changed employers in year 3 contribute 3 person-years. Censored individuals—those who were still with the same employers in year 5—contribute the maximum of 5 person-years. For the 200 biochemists, there were a total of 848 person-years. From Table 1 it can be seen that this total is just the sum of the number at risk in each of the five years.

[p. 19 ↓] For each person-year, the dependent variable is coded 1 if a person changed employers in that year, otherwise it is coded 0. The explanatory variables are assigned the values they took on in each person-year.[1] The final step is to pool the 848 person years into a single sample, and then estimate logit models for a dichotomous dependent variable using the method of maximum likelihood. Programs for maximum likelihood logit analysis are now widely available, for example, in the SAS (SAS Institute, 1983), BMDP (Dixon, 1981), SPSSX (SPSS Inc., 1984), or GLIM (Baker and Nelder, 1978) statistical packages.

Notice how the two problems of censoring and time-varying explanatory variables are solved by this procedure. Individuals whose time to the first job change is censored contribute exactly what is known about them, namely that they did not change jobs in any of the five years in which they were observed. Time-varying explanatory variables are easily included because each year at risk is treated as a distinct observation.

Estimates for the Biochemistry Example

Let us see what happens when this procedure is applied to the biochemistry data. Table 2 reports estimates for Model 1 which does not allow the hazard rate to vary autonomously with time. The coefficient estimates are like unstandardized regression coefficients in that they depend on the metric of each independent variable. For our purposes, it is instructive to focus on the t-statistics for the null hypothesis that each

coefficient is zero. (The column labeled OLS t will be discussed later.) These are metric-free and give some indication of the relative importance of the variables.

Three of the variables have a significant impact on the hazard rate for changing employers. Specifically, biochemists with many citations are more likely to change employers, while associate professors and those employed at institutions receiving high levels of funding are less likely to change employers. (These results suggest that most of the job changes are voluntary.) Prestige of department and number of publications seem to make little difference.

Model 2 allows the hazard rate to be different in each of the five years, even when other variables are held constant. This was accomplished by creating a set of four dummy variables, one for each of the first four years of observation. Coefficient estimates and test statistics are shown in Table 2. The coefficient for each dummy variable gives the difference in the logit of changing employers in that year and the logit of changing employers in year 5, holding other variables constant. No clear pattern emerges from these coefficients, although there does appear to be some**[p. 20 ↓]** tendency for the hazard rate to increase with time. In this example, the introduction of the dummy variables makes little difference in the estimated effects of the other variables, but this will not always be the case.

TABLE 2 Estimates for Logit Models Predicting the Probability of an Employer Change, 848 Person-Years

TABLE 2
Estimates for Logit Models Predicting the Probability of an Employer Change, 848 Person-Years

Explanatory Variables	*Model 1*			*Model 2*		
	b	*t*	*OLS t*	*b*	*t*	*OLS t*
Prestige of Department	.045	.21	.22	.056	.26	.25
Funding	−.077	−2.45*	−2.34	−.078	−2.47*	−2.36
Publications	−.021	−.75	−.86	−.023	−.79	−.91
Citations	.0072	2.44*	2.36	.0069	2.33*	2.23
Rank (D)[a]	−1.4	−2.86**	−2.98	−1.6	−3.12**	−3.26
Year 1 (D)				−.96	−2.11*	−2.07
Year 2 (D)				−.025	−.06	.18
Year 3 (D)				−.74	−1.60	−1.54
Year 4 (D)				−.18	−.42	−.38
Constant	4.95			2.35		
Log-likelihood	−230.95			−226.25		

a. (D) indicates dummy variable.
*Significant at .05 level, 2-tailed test.
**Significant at .01 level, 2-tailed test.

The Likelihood-Ratio Chi-Square Test

By comparing Models 1 and 2, one can test the null hypothesis that the hazard rate for changing employers does not vary with time, net of other variables. The procedure is very similar to testing for the significance of increments to R^2 resulting from the addition of explanatory variables to a multiple regression equation. The test is applicable whenever one model is a special case of another. This occurs, for example, if one model includes all the variables that are in another model, but also includes additional variables. The test statistic is constructed from a byproduct of maximum likelihood estimation, the maximized value of the log-likelihood function. This is given in Table 2 for each of the two models. To compare the fit of two models, one calculates twice the positive difference between their log-likelihoods. (Instead of the log-likelihood, some computer programs report -2 times the log-likelihood in order to facilitate computation of this statistic.) Under the null hypothesis of no difference, this statistic will have an asymptotic chisquare**[p. 21 ↓]** distribution. The associated degrees of freedom will be the number of constraints that distinguish the two models; in most cases, this will just be the difference between the numbers of variables in the two models. In this example, twice the difference in the log-likelihoods is 9.4. Since Model 1 has four fewer variables than Model 2, there are four degrees of freedom. This chi-square value is just below the critical value for the .05 level of significance. Thus the evidence is marginal that the hazard rate varies autonomously with time.[2]

This procedure for comparing log-likelihoods to test hypotheses about *sets* of variables is quite generally applicable for maximum likelihood estimation. It can therefore be applied to any of the models and estimation procedures to be discussed in later chapters.

Problems with the Discrete-Time Method

In this example, the number of constructed person-years was quite manageable with respect to computation. On the other hand, when a large sample is followed over a long interval divided into small discrete units of time, the resulting number of constructed

observations can be impractically large. In the biochemistry example, switching to person-months instead of person-years would produce a working sample of nearly 10,000. One can always solve the problem by aggregating data into larger intervals of time, but that necessarily discards some information. (Note, however, that little information is lost in the biochemistry example because most academic job changes occur at the beginning of the academic year.)

Allison (1982) discusses several ways in which discrete-time methods can be applied to large samples with minimal cost. For example, if all the explanatory variables are categorical (nominal), estimation of the logit model can be done by log-linear methods. It is well-known that with log-linear models the computational cost depends on the number of cells in the contingency table, not the number of observations in the cells.

Another way to reduce the cost of the discrete-time method is to do exploratory analysis using ordinary least squares (OLS) multiple regression on the pooled set of individual time units to estimate the linear probability model in equation 1. OLS is much cheaper because, unlike maximum likelihood logit analysis, it is not an iterative method, and, once the correlation matrix is constructed, alternative models can be estimated at extremely low cost. As an example, OLS regressions were performed on the 848 biochemistry person-years with the dependent variable coded 1 if an employer change occurred and otherwise coded**[p. 22 ↓]** zero. The t-statistics for these regressions are given in Table 2, next to the t-statistics for the logit model. The results are remarkably similar. (For guidance as to when least squares will give a good approximation, see Goodman, 1975.)

Discrete Versus Continuous Time

Before moving on to continuous-time methods, it must be stressed that the discrete-time method described here will virtually always give results that are quite similar to the continuous-time methods described later. In fact, as the time units get smaller and smaller, the discrete-time model of equation 3 converges to the proportional hazards model of Chapter 4. While there is some loss of information that comes from not knowing the exact time of the event, this loss will usually make little difference in the estimated standard errors.

Thus, the choice between discrete- and continuous-time methods should generally be made on the basis of computational cost and convenience. When there are no time-varying explanatory variables, it is usually simpler to do event-history analysis using one of the methods described in the next two chapters. This is largely due to the fact that the continuous-time methods do not require that the observation period for each individual be subdivided into a set of distinct observational units. When there are time-varying explanatory variables, on the other hand, the relative costs and convenience of using continuous-time and discrete-time methods are quite comparable.

http://dx.doi.org/10.4135/9781412984195.n2