Event History Analysis

PARAMETRIC METHODS FOR CONTINUOUS-TIME DATA

Contributors: Paul D. Allison
Book Title: Event History Analysis
Chapter Title: "PARAMETRIC METHODS FOR CONTINUOUS-TIME DATA"
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.n3
Print pages: 23-34

http://dx.doi.org/10.4135/9781412984195.n3

PARAMETRIC METHODS FOR CONTINUOUS-TIME DATA

Although the discrete-time method just discussed is widely applicable, most event history analysis is done using continuous-time methods. In this chapter we shall examine some of the more popular parametric methods for data in which time is measured precisely. These methods are called parametric because every aspect of the model is completely specified, except for the values of certain parameters, which must be estimated. Attention is restricted to situations in which each individual experiences no more than one event and all events are treated alike.

There are many closely related approaches to the analysis of such data, and the novice may be hard pressed to choose among them. We shall try to provide some guidelines for that choice. While a deep**[p. 23 ↓]** understanding of these methods requires knowledge of calculus (including simple ordinary differential equations) and maximum likelihood, it is possible to become an intelligent user with only a modest mathematical background.

The Continuous-Time Hazard Rate

What nearly all these methods share is the notion of the hazard rate as the fundamental dependent variable. In the previous chapter, the discrete-time hazard was defined as the probability that an individual experiences an event at time t, given that the individual was at risk at time t. This definition will not work in continuous time, however, because the probability that an event occurs at *exactly* time t is infinitesimal for every t. Instead, consider the probability that an individual experiences an event in the *interval* from t to t + s, given that the individual was at risk at time t, and denote this probability by P(t, t + s). When s = 1, this is equivalent to the discrete-time hazard defined in Chapter 2. Next we divide this probability by s, the length of the interval, and let s become smaller and smaller until the ratio reaches a limit. This limit is the continuous-time hazard, denoted by h(t). Other common symbols for the hazard rate are #(t) and r(t). Formally,

$$h(t) = \lim_{s \to 0} P(t, t+s)/s \qquad [4]$$

Although it may be helpful to think of this as the instantaneous probability of event occurrence, it is not really a probability because it can be greater than 1. In fact, it has no upper bound. A more accurate interpretation is to say that h(t) is the unobserved rate at which events occur. Specifically, if h(t) is constant over time, say h(t) = 1.25, then 1.25 is the expected number of events in a time interval that is one unit long. Alternatively, 1 / h(t) gives the expected length of time until an event occurs, in this case .80 time units. This way of defining the hazard corresponds closely to intuitive notions of risk. For example, if two persons have hazards of .5 and 1.5, it is appropriate to say that the second person's risk of an event is three times greater.

For most applications, it is reasonable to assume that the hazard rate changes as a function of time, either the time since the last event or the age of the individual. For example, available evidence indicates that, at least after age 25, the hazard rate for being arrested declines with age. **[p. 24 ↓]** The hazard for death from any cause has a U shape: It is relatively high immediately after birth, declines rapidly in the early years, and then begins to rise again during late middle age.

It is important to realize that the shape of the hazard rate function is one of the key distinguishing features of different models for continuous-time data. In fact, the hazard function h(t) completely determines the probability distribution of the time until an event (or the time between events when events are repeata#able). Later in this chapter we shall see how one might go about choosing a shape for the hazard function.

Continuous-Time Regression Models

The next step is to develop models for the dependence of h(t) on time and on the explanatory variables. We shall consider three models—the exponential, the Weibull, and the Gompertz—that differ only in the way that time enters the equation. To keep it simple, let us assume that we have only two explanatory variables, x
1

and x
2

, which do not vary over time. An obvious approach would be to let h(t) be a linear function of the explanatory variables. This is awkward, however, because h(t) cannot be less than zero, and there is nothing to prevent a linear function from being less than zero. It is typical, then, to take the natural logarithm of h(t) before setting it equal to a linear function of the explanatory variables. Thus, one of the simplest models is

$$\log h(t) = a + b_1x_1 + b_2x_2 \qquad [5]$$

where a, b
1

, and b
2

are constants to be estimated. In this equation h(t) is a function of the explanatory variables, but it does not depend on time. A hazard that is constant over time implies an exponential distribution for the time until event occurrence and, hence, this is often referred to as the exponential regression model.

Specifying a constant hazard is usually unrealistic, however. If the event is a death, for example, the hazard should increase with time, due to aging of the organism. On the other hand, if the event is an employer change, the hazard is likely to decline with time as the individual becomes more invested in the job. We can relax the assumption of a constant hazard by allowing the log of the hazard to increase or decrease linearly with time, i.e.,

$$\log h(t) = a + b_1x_1 + b_2x_2 + ct \qquad [6]$$

[p. 25 ↓] where c is a constant which may be either positive or negative. Because this model gives rise to a Gompertz distribution for the time until event occurrence, it is convenient to refer to equation 6 as the Gompertz regression model.

Alternatively, let us consider a model in which the log of the hazard increases or decreases linearly with the log of time:

$$\log h(t) = a + b_1x_1 + b_2x_2 + c \log t \qquad [7]$$

where c is constrained to be greater than -1. This model generates a Weibull distribution for the time until event occurrence. Hence, it is often referred to as a Weibull regression model.

There are many other models that differ only in the way that time enters the equation, but these three are the most common. For additional information on these three models, see Lawless (1982). Although time appears to be just another explanatory variable in the Weibull and Gompertz models, its role is much more fundamental. In particular, the difference between equations 6 and 7 requires an entirely different estimation procedure rather than a simple transformation from time to log time.

Note that neither the Weibull model nor the Gompertz model allows for a U shape or an inverted U shape; the hazard may either decrease or increase with time, but may not change direction. This is a disadvantage in some applications. Later we shall consider some models that do not have this restriction.

Notice also that none of these models has a random disturbance term. They are not deterministic models, however, because there is random variation in the relationship between the unobservable dependent variable h(t) and the observed length of the time interval. Still, there are some who argue that these models should include a disturbance term, an issue that will be discussed at the end of this chapter.

Maximum Likelihood Estimation

Writing down models is easy. The difficulty comes in trying to estimate them, especially with censored data. In the late 1960s, statisticians developed maximum likelihood procedures for the exponential regression model (Zippin and Armitage, 1966; Glasser, 1967), and it was not long before maximum likelihood was available for many other

models as well. Appendix A discusses maximum likelihood estimation of parametric models in some detail, but it is worth mentioning some of the general properties here.

[p. 26 ↓] As an estimation method for censored data, maximum likelihood is hard to beat. It combines the censored and uncensored observations in such a way as to produce estimates that are asymptotically unbiased, normally distributed and efficient (i.e., have minimum sampling variance). Unfortunately, “asymptotically” means that these properties are only approximations that improve as the sample gets larger. No one knows how well they hold in small samples or how large is large enough. In the absence of compelling alternative methods, however, maximum likelihood is widely used with both small and large samples.

There are many computer programs available to do maximum likelihood estimation of one or more of these models. The RATE program (Tuma, 1979) will estimate the exponential model, the Gompertz model, and several extensions of the Gompertz model. The GLIM program (Baker and Nelder, 1978) will estimate the exponential and Weibull regression models (in version 3), but only by employing special procedures that are not documented in the manual (Aitkin and Clayton, 1980; Roger and Peacock, 1983). Weibull and exponential models can also be estimated with two author-distributed programs, CENSOR (Meeker and Duke, 1981) and SURVREG (Preston and Clarkson, 1983).

An Empirical Example

To illustrate these methods, we shall apply the exponential regression model to the criminal recidivism data (Rossi et al., 1980) that were briefly described in Chapter 1. The sample consisted of 432 males who were followed for one year after their release from Maryland state prisons. The study was actually a randomized field experiment in which approximately half the men received financial assistance while the other half served as a control group. During the follow-up year, the subjects were interviewed monthly regarding their experiences during the previous month. At the end of the year, a search was made through district court records for data on arrests and convictions.

The event of interest is the first arrest after release, and the aim is to determine how the hazard for an arrest depends on the following explanatory variables: age, race, years of schooling, marital status, age at earliest known arrest, number of previous theft arrests, parole status, financial assistance, prior work experience, and number of weeks employed during the first three months after release. With the exception of the last variable, all of these variables are clearly constant in value over the follow-up period. While employment status is obviously changeable**[p. 27 ↓]** over the full year after release, this analysis will treat it as constant over time. Later, we shall examine a model allowing employment status to vary over time.

TABLE 3 Coefficient Estimates for Three Models of Recidivism

TABLE 3
Coefficient Estimates for Three Models of Recidivism

	1		*2*		*3*	
Explanatory Variables	*Exponential*		*Proportional Hazards*		*Time-Dependent Proportional Hazards*	
	b	*t*	*b*	*t*	*b*	*t*
Financial aid (D)[a]	−.325	−1.69	−.337	−1.76	−.333	−1.74
Age at release	−.067	−2.89**	−.069	−2.94**	−.064	−2.78**
Black (D)	.280	.90	.286	.92	.354	1.13
Work experience (D)	−.117	−.53	−.122	−.55	−.012	−.06
Married (D)	−.414	−1.08	−.426	−1.11	−.334	−.87
Paroled (D)	−.037	−.19	−.035	−.18	−.075	−.38
Prior arrests	.095	3.21**	.101	3.36**	.100	3.31**
Age at earliest arrest	.070	2.30*	.071	2.35*	.077	2.48*
Education	−.263	−1.96*	−.264	−1.96*	−.293	−2.12*
Weeks worked	−.039	−1.76	−.039	−1.78	–	–
Worked (D)	–	–	–	–	−1.397	−5.65**
Constant	−3.860		–		–	

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

Most programs for estimating event history models require that the data on the dependent variable be input in two parts: a dummy variable indicating whether or not the event (in this case an arrest) occurred during the observation period, and a variable giving either the time of the event (if it occurred) or the time of censoring. In this example, time was measured in weeks since release. Thus, for those who were arrested, the second component of the dependent variable was the number of weeks from release to arrest. For those who were not arrested, the week number was 52, the last week that they were observed. Estimates for an exponential regression model were obtained with the GLIM program (see Appendix B for program listing), and are reported in panel 1 of Table 3.

Interpreting the coefficient estimates is much like interpreting unstandardized regression coefficients. For example, the coefficient of -.067 for age at release means that

each additional year of life reduces the log of the hazard by .067, controlling for other variables. A somewhat**[p. 28 ↓]** more intuitive interpretation is obtained by exponentiating the coefficients (taking their antilogs). That is, if b is the coefficient, compute exp(b), which means raising the number e (approximately 2.718) to the b power. The interpretation is then as follows: For each unit increase in an explanatory variable, the hazard is multiplied by its exponentiated coefficient. Further, computing 100(exp(b) - 1) gives the percentage change in the hazard with each one unit change in the explanatory variable. For example, the coefficient for number of prior arrests is .095, and exp(.095) = 1.10. This tells us that each additional prior arrest increases the hazard by an estimated 10 percent. For dummy variables, the exponentiated coefficient gives the relative hazard for the groups corresponding to values of the dummy variable, again controlling for other variables. The coefficient for the dummy variable for financial aid, for instance, is -.325, which gives exp(-.325) = .72. This means that the hazard of arrest for those who received financial aid was about 72 percent of the hazard for those who did not recieve aid. Alternatively, since 1/.72 = 1.38, we can say that the hazard for those who did not receive aid was about 38 percent larger than the hazard for those who did receive aid.

The ratios of the estimates to their standard errors are also useful statistics. For moderate to large samples, these can be treated like t-statistics in an ordinary multiple regression. Thus if the ratio exceeds 2, the coefficient is significantly different from zero at the .05 level with a two-tailed test. Also the relative sizes of these ratios can be used to gauge the relative importance of the variables. In this example, we see that only four of the 10 variables have effects which are definitely significant: age at release, age at earliest known arrest, education, and number of prior arrests. The effect of financial aid is significant with a one-tailed test but not with a two-tailed test. With the exception of age at first arrest, all these effects are in the expected direction. Thus, the positive sign of prior arrests means that those with many prior arrests have a higher risk of being arrested at any point in time.[3]

Censoring

In both this example and the biochemistry example in Chapter 2, censoring occurred at the same point in time for all individuals. Thus, any released inmates who had still not been arrested at the end of 12 months were considered censored. This is sometimes called fixed censoring or Type I censoring. Under fixed censoring, it is unnecessary to make any further assumptions about the nature of the censoring process.

In many situations, however, the censoring times will vary across individuals. This occurs when individuals drop out of the study, for one**[p. 29 ↓]** reason or another, before the end of the observation period. Possible reasons include death, migration out of the population at risk, failure to locate the individual in later interviews, or refusal to continue in the study. When censoring times vary across individuals (and are not under the control of the investigator) censoring is said to be random. Random censoring also includes designs in which observation *ends* at the same time for all individuals, but begins at different times.

When censoring is random, virtually all event history methods assume that the censoring times are independent of the times at which events occur, controlling for the explanatory variables in the regression model. This assumption would be violated, for example, if individuals who were more likely to be arrested were also more likely to migrate out of the study area. Although it is possible to develop models which allow for dependence between censoring and the occurrence of the event of interest, this is rarely done. The main reason why it is not often done (aside from the inconvenience of a nonstandard model) is that it is impossible to test whether any dependence model is more appropriate than the independence model (Tsiatis, 1975). In other words, the data can never tell you which is the correct model.

It is possible, however, to get some idea of how sensitive one's analysis is to violations of the independence assumption. In essence, the sensitivity analysis consists of reestimating the model twice, each time treating the censored observations in a different extreme way. The first step is to redo the analysis with the data altered so that censored observations experience an event at the time of censoring. In most cases, this is easily accomplished by recoding the dummy variable which indicates whether or not an

observation is censored so that all observations have a value of 1. The second step is to redo the analysis so that the censoring times are all equal to the longest time observed in the sample, regardless of whether that time is censored or uncensored. Thus, even if some of the released prisoners had been censored before the end of the one-year period, their censoring times would be set to one year. If the parameter estimates resulting from the standard analysis are similar to those obtained from these two extreme situations, one can be confident that violations of the independence assumption are unimportant (Peterson, 1976). Note that this approach to censoring can be used with any of the methods discussed in later chapters.

Some Other Models

The three models considered above—the exponential, the Weibull, and the Gompertz —are all members of a general class of models known as proportional hazards models. In the next chapter, we shall see how to**[p. 30 ↓]** estimate this general class without having to choose a particular member. Before leaving the subject of parametric models, however, let us briefly consider another general class of models known either as accelerated failure time models (Kalbfleisch and Prentice, 1980) or as location-scale models (Lawless, 1982). If T is the elapsed time until an event occurs, this class can be written as

$$\log T = a + b_1x_1 + b_2x_2 + \ldots + u \qquad [8]$$

where u is a random disturbance term that is independent of the x's.

Different members of this class have different distributions for the disturbance term u. Distributions that are commonly assumed include the normal, log-gamma, logistic, and extreme value distributions. These give rise, respectively, to log-normal, gamma, log-logistic, and Weibull distributions for T. Thus, the Weibull regression model is a member of both the proportional hazards class and the accelerated failure time class. In fact, it can be shown that the Weibull (and its special case—the exponential) is the only model that falls into both of these classes.

The accelerated failure time models can be reexpressed so that the dependent variable is the hazard rate rather than log T, but these expressions tend to be quite complicated. The lognormal and log-logistic models are unusual in that the hazard is a nonmonotonic function of time; it first increases, reaches a peak, and then gradually declines with time. When there is no censoring of T, the accelerated failure time models can be consistently estimated by ordinary least squares regression of log T on the explanatory variables. In the presence of censoring, however, one must usually resort to maximum likelihood estimation. For details see Lawless (1982).

Choosing a Model

How does one choose among alternative parametric models? As with most statistical methods, it is rather difficult to codify the procedures involved in choice of a model. There are many factors that should legitimately enter the decision and none can be easily quantified. Invariably there is tension among mathematical convenience, theoretical appropriateness, and empirical evidence.

With models of the sort we have just been discussing, the key differentiating factor is the way in which the hazard rate depends on time. The first choice is between the exponential regression model, in which there is no dependence on time, and all other models. From a mathematical and computational point of view, the exponential model is very attractive.**[p. 31 ↓]** For this reason, it is often useful as a first approximation even when it is known to be false. Substantive theory, on the other hand, will usually suggest many reasons why the hazard should change with time. As for empirical evidence, there are well-known graphical methods for assessing whether event times have an exponential distribution (Gross and Clark, 1975; Elandt-Johnson and Johnson, 1980, Ch. 7; Lawless, 1982, Ch. 2). These can be quite useful if the explanatory variables have relatively weak effects on the hazard. If effects are strong, however, the graphical methods may show a declining hazard even when the true hazard is constant over time.

A better approach is to fit the exponential regression model with explanatory variables and then examine its fit to the data. This can be done by using the residual plots described by Lawless (1982) for evidence of departure from exponentiality. A more formal method for testing the fit of the exponential regression model derives from the

fact that the exponential is a special case of several of the models considered thus far, including the Weibull, Gompertz, and gamma regression models. The procedure is to fit both the exponential and one of these other models by maximum likelihood. The relative fit of the two models can then be tested by comparing log-likelihoods as described in Chapter 2. Rejection of the exponential model is indicated when its log-likelihood differs significantly from that of the alternative model.

If the exponential model is rejected, one must then choose between monotonic models (in which the hazard always increases or always decreases with time) and nonmonotonic models (in which the hazard may sometimes increase and sometimes decrease). Again, both substantive theory and the graphical methods referenced above can be helpful in making this choice. As before, however, one must be wary of univariate graphs for assessing the shape of the hazard function. Strong effects of explanatory variables can make the evidence misleading. As an alternative to univariate graphical techniques, residual plots are available for assessing the fit of a chosen model (Lawless, 1982).

As noted above, the lognormal and log-logistic models have nonmonotonic hazard functions in which the hazard first increases and then decreases. This might be appropriate for many kinds of social mobility in which there is (a) an initial “resting period” after the previous move, (b) an increase in the risk of a move as the resting period is completed, and (c) a decline in the risk of a move as individuals become more invested in a particular social location. On the other hand, there is no convenient parametric model to represent U-shaped hazard functions. If there is strong departure from monotonicity, it is often better to shift**[p. 32 ↓]** to the semiparametric, proportional hazards model discussed in the next chapter.

Within the class of monotonic models, choice of model will often be based on mathematical and computational convenience. Social theory and empirical evidence are typically inadequate to discriminate between, say, a Weibull model and a Gompertz model.[4]

Unobserved Sources of Heterogeneity

Many social theories imply or suggest that the hazard rate for some event should be increasing or decreasing with time. For example, certain job search theories imply that the hazard rate for obtaining a job should increase with the length of unemployment (Heckman and Singer, 1982). While the procedures described in the preceding section on model choice can be useful in testing such hypotheses, great caution is required in trying to draw inferences about the effect of time on the hazard rate. The basic problem was mentioned above, but is worth some elaboration here. Even if the hazard rate is constant over time for each individual, differences (across individuals) in the hazard rate that are not incorporated into the model will tend to produce evidence for a *declining* hazard rate (Heckman and Singer, 1982).

Intuitively, what happens is that individuals with high hazard rates experience events early and are then eliminated from the risk set. As time goes on, this selection process yields risk sets that contain individuals with predominantly low risks. The upshot is that it is extremely difficult to distinguish hazard rates that are truly declining with time from simple variation in hazard rates across individuals. On the other hand, if one observes evidence for an *increasing* hazard rate, this can always be regarded as evidence that the hazard really increases with time.

A simple way to deal with the problem of heterogeneity is to explicitly incorporate the sources of that heterogeneity as explanatory variables in the model. But it would be unrealistic to assume that all such sources of heterogeneity can be measured and included. This problem has been a matter of concern for sociologists and econometricians working with these models, and there have been attempts to expand the model to include a disturbance term representing the unobserved sources of heterogeneity. In particular, Heckman and Singer (1982) have considered an extended Weibull model,

$$\log h(t) = a + b_1x_1 + b_2x_2 + c \log t + u \qquad [9]$$

[p. 33 ↓] where u is a random disturbance. In principle, estimation of such a model should allow one to separate the effects of time from the unobserved heterogeneity. In practice, they have found that estimates of c and the b coefficients are highly sensitive to the choice of a particular distribution for u. Although work is being done to remedy this problem, it is still too early to conclude whether the approach will be generally useful.

Biostatisticians have been remarkably unconcerned about the problem of unmeasured sources of heterogeneity, though there is every reason to suspect that it will be just as serious for biological as for social phenomena. Their attitude seems to be that the consequence of such heterogeneity will be mainly to change the shape of the distribution of T (the time of the event), and that this can be accommodated by specifying a different distribution for T (e.g., a Gompertz instead of a Weibull), or by using a more general model (e.g., the proportional hazards model considered in the next chapter). This position is reasonable so long as one is primarily concerned with estimating the effects of the explanatory variables and is not particularly interested in testing hypotheses about the effect of time.

http://dx.doi.org/10.4135/9781412984195.n3