30
Cox's method for follow-up studies

When using Poisson regression models to analyse data from follow-up studies, time is divided into fairly broad bands such as 5 or 10 years of age. Age is the most common time scale but in some applications other time scales may be more relevant. This point is discussed in more detail in the next chapter, but for the moment we refer to the time scale simply as time. Cox's method is very similar to Poisson regression but is based on a much finer subdivision of time.

30.1 Choosing parameters

When there are two explanatory variables, A and B, and the rate is allowed to vary with time, the multiplicative model for the rate takes the form

$$\text{Rate} = \text{Corner} \times \text{Time} \times \text{A} \times \text{B}.$$

Here time is a categorical variable with one level for each time band. Again we split the model into two parts, as in

$$\text{Rate} = \boxed{\text{Corner} \times \text{Time}} \times \boxed{\text{A} \times \text{B}}\,.$$

Algebraically this corresponds to a reparametrization of the model as

$$\lambda_i^t = \lambda_{\text{C}}^t \theta_i,$$

where λ_{C}^t is a corner parameter measuring the rate for time band t when A and B are both at level 0, and θ_i is the rate ratio which compares the rate for subject i, in time band t, to the corner rate for that time band. The parameters λ_{C}^t correspond to the

$$\boxed{\text{Corner} \times \text{Time}}$$

part of the model and the parameters θ_i to the

$$\boxed{\text{A} \times \text{B}}$$

part of the model.

30.2 The profile likelihood

The parameters λ_{C}^t are also called the *baseline* rates, and are generally nuisance parameters. The main interest is in the parameters of the second part of the model. The profile likelihood for the parameters in the second part of the model is obtained by deriving formulae for the most likely values of the nuisance parameters, λ_{C}^t, and substituting these into the expression for the log likelihood. The number of nuisance parameters depends upon the number of time bands into which the total study period has been partitioned. For the present we shall consider a finite number of bands, but in the next section the argument is generalized to the case where time is divided into clicks.

The contribution of subject i to the log likelihood is the sum of contributions for each time band. These have the Poisson form:

$$d_i^t \log(\lambda_i^t) - y_i^t \lambda_i^t$$

where y_i^t is the observation time in time-band t and d_i^t indicates whether the event occurred ($d = 1$) or not ($d = 0$). The total log likelihood is the sum of such terms over all subjects (i) and all time bands (t). Rewriting λ_i^t as $\lambda_{\text{C}}^t \theta_i$, this becomes

$$\sum_{i,t} \left[d_i^t \log(\lambda_{\text{C}}^t \theta_i) - y_i^t \lambda_{\text{C}}^t \theta_i \right].$$

The rules of calculus show that, given the θ_i, the most likely values of the baseline rates λ_{C}^t are

$$\frac{d^t}{\sum_i y_i^t \theta_i},$$

where d^t represents the total number of events occurring in time band t. Substituting these values into the expression for the log likelihood yields a profile log likelihood which depends only on the parameters in the second part of the model. This is

$$\sum_{j,t} d_j^t \log\left(\frac{\theta_j}{\sum_i y_i^t \theta_i} \right).$$

30.3 Time divided into clicks

The profile log likelihood derived by stratifying the follow-up interval into bands provides a satisfactory method for regression analysis of cohort studies, but although this is the approach used with frequency records it is rarely used with individual records. The reason for this is that a further generalization offers increased flexibility without seriously compromising either statistical or computational efficiency. In this generalization the time scale is subdivided into clicks which can contain no more than one event, thus allowing rates to vary continuously over time.

The consequence of this generalization for the profile log likelihood are quite minor. First consider the effect upon the observation times, y_i^t. If the duration of the time bands is h and we allow h to become very small, almost every y_i^t will become either zero (if subject i was not observed at click t) or h (if subject i was observed). In these circumstances, it is convenient to redefine y_i^t to be *at risk indicators* taking on the values 0 or 1 respectively. The observation times then become hy_i^t and the profile log likelihood for the rate ratio model becomes

$$\sum_{j,t} d_j^t \log\left(\frac{\theta_j}{\sum_i hy_i^t\theta_i}\right),$$

which may be further simplified to

$$\sum_{j,t} d_j^t \log\left(\frac{\theta_j}{\sum_i y_i^t\theta_i}\right) - D\log(h).$$

Since the term $D\log(h)$ does not depend upon any parameters, it may be omitted.

Examination of the profile likelihood equation shows it to be constructed of a sum of terms, in which d_j^t is a multiplier which takes on the value 1 for clicks in which an event occurs, and 0 everywhere else. Thus the profile log likelihood receives an additive contribution for every failure event. Each of these is the log of a ratio whose numerator is the rate ratio, θ_j, predicted by the model for subject j in whom the event occurred (the *case*), and whose denominator,

$$\sum_i y_i^t\theta_i$$

is the sum of rate ratios, θ_i, for those subjects under observation at t, the time of occurrence of the failure.

The collection of subjects contributing to the denominator is known as the *risk set* for the observed failure. Using this terminology the profile likelihood can be written

$$\sum_{\text{Failures}} \log\left(\theta_{(\text{for case})} \Big/ \sum_{\text{Risk set}} \theta\right).$$

The ratio in brackets is the conditional probability that, given a failure occurred in this set of subjects, it occurred in the case rather than in some other member of the risk set. The profile log likelihood therefore

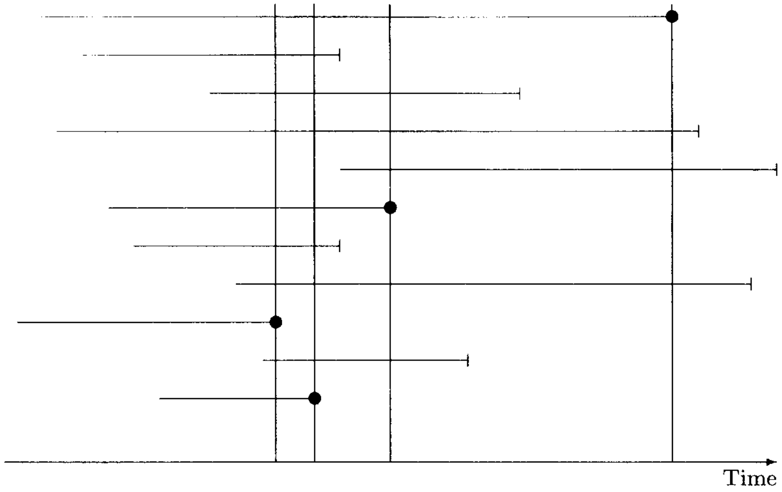

Fig. 30.1. Composition of risk sets.

corresponds exactly with the conditional log likelihood obtained for individually matched case-control studies, and analysis of a cohort study using the above profile likelihood is equivalent to its analysis as a matched case-control study in which each case is matched on time with all other members of the corresponding risk set. The composition of risk sets is illustrated by Fig. 30.1. The risk set for each failure contains all subjects whose observation lines cross the appropriate vertical, including the subject in whom the defining event occurred.

The recognition that this likelihood is a profile likelihood came some years after Cox's original proposal of the method, in which he called it the *partial likelihood.*[*] This name has stuck, and is in general use, so we shall continue to use it, but we emphasize that partial likelihood is the profile likelihood for the parameters in the second part of the regression model when Cox's method has been used to eliminate the parameters in the first half. Because a very large number of nuisance parameters have been eliminated — infinitely many, in fact, we have no right to expect that the partial likelihood will maintain the properties of likelihood. In the present application, however, it has been proved to behave the same way as a true

[*] Cox originally used an argument identical to that we used in Chapter 29 for individually matched case-control studies and referred to it as a *conditional* likelihood. There are, however, difficulties with this argument when applied in the present context. While each term which contributes to the log likelihood is indeed the logarithm of a conditional probability, the total is not. A later paper correcting this error introduced the term partial likelihood.

Table 30.1. A cohort of 10 subjects

		Entry to Study		End of Study	
Subject	Sex	Date	Age	Date	Age
A	F	13/ 6/65	29.3	31/12/89	53.8
B	M	23/10/72	25.2	31/12/89	42.4
C	M	3/ 3/59	22.1	31/12/89	52.8
D	F	10/10/67	32.2	31/12/89	54.4
E	M	2/ 1/60	33.1	4/ 7/79	52.6
F	M	9/ 1/75	42.1	31/12/89	57.1
G	F	5/ 8/53	35.2	3/10/68	50.4
H	M	10/10/69	27.0	31/12/89	47.2
I	M	2/ 3/72	44.8	31/12/89	62.7
J	F	1/11/70	51.5	31/12/89	70.6

likelihood as the amount of data increases.

The composition of risk sets (and hence the results of the analysis) depend upon the choice of time scale for the analysis, as is demonstrated by the following exercise.

Exercise 30.1. The data set out in Table 30.1 refer to 10 subjects from a cohort study. Subjects E and G died at the second date while the remaining eight subjects survived until the date of analysis (31/12/89). List the members of the risk sets for both deaths when the appropriate time scale is (a) calendar date (b) age (c) time since entry into the study.

The difference between these analyses is that they represent three different models. In each case the λ_C^t parameters represent variation of baseline rates along different time scales.

30.4 Choice of time scale

Our derivation of Cox's method allows for time to be interpreted in the most appropriate manner for a particular analysis. Usually this will mean the time scale with the strongest relationship to failure rate. Regrettably it is still the case that some major software packages do not allow such flexibility. This reflects the fact that the method was motivated by problems of survival following medical treatment. In such studies the appropriate time scale is time since start of follow-up so that all observation of all subjects starts at time zero. In such studies, risk sets always become smaller (as a result of failure and censoring) as time advances.

On other time scales there will be *late entry* of subjects (observation starting at time > 0) and risk sets may be supplemented by new entrants as time advances. In order to be able to select the most appropriate time scale for an analysis, the software must be capable of allowing for late entry.

30.5 Confounders other than time

The confounding effect of time is allowed for by including time in the first part of the model. For example, taking age as the time variable, the multiplicative model

$$\text{Rate} = \boxed{\text{Corner} \times \text{Age}} \times \boxed{\text{A} \times \text{B}},$$

includes the effect of age in the baseline rate parameters. The most obvious way to deal with another confounder, such as sex, is to include it in the second part of the model, as in

$$\text{Rate} = \boxed{\text{Corner} \times \text{Age}} \times \boxed{\text{Sex} \times \text{A} \times \text{B}}.$$

This model assumes that the effect of sex is constant with age so that the baseline rates for males are a constant multiple of those for females. To extend the model to allow for different patterns of baseline rates for each sex, the interaction between age and sex must be included in the model. When the age scale is divided into clicks this interaction term involves a very large number of parameters, so it is best to absorb these parameters in the baseline rate part of the model, giving

$$\text{Rate} = \boxed{\text{Corner} \times \text{Age} \times \text{Sex} \times \text{Age}{\cdot}\text{Sex}} \times \boxed{\text{A} \times \text{B}}.$$

This model has the effect of allowing different sets of baseline rate parameters for males and females. If we estimate these algebraically as before, we find that the profile likelihood for the rate ratio part of the model still has the form of a partial likelihood:

$$\sum_{\text{Failures}} \log \left(\theta_{\text{(for case)}} \Big/ \sum_{\text{Risk set}} \theta \right)$$

but the risk set is now restricted to contain only those subjects who (a) were under study at the time of failure of the case, and (b) belonged to the same sex as the case. Thus the analysis simulates a matched case-control study in which controls are matched to cases with respect to sex.

This extension of Cox's method is usually referred to as a stratified analysis, although more properly it should be referred to as *doubly* stratified — Cox's method stratifies by time alone, while the extended method stratifies by both time and a further variable. In our example stratification is by age and sex.

Exercise 30.2. Repeat Exercise 30.1 for an analysis which is to be stratified by sex.

It can be seen from the last exercise that when an analysis is doubly stratified the risk sets contain fewer subjects than when it is stratified on time alone. Rather unexpectedly, therefore, the effect of adopting a more complicated model is to *reduce* the amount of computation required to estimate the parameters of interest. Further stratification can be introduced but there is a limit. If a study is overstratified, some risk sets will contain only the case, there being no other subjects matching the case in respect of all stratifying variables. Such sets make no contribution to the profile likelihood, so the information from these events is lost.

★ 30.6 Estimating the baseline rates

In some circumstances the dependence of rates upon time is of some interest, and we would wish to estimate the baseline rates, λ^t_{C}. In this section we shall show that the plot of the most likely estimate of the baseline rate against time turns out to be very similar in form to the Aalen– Nelson estimator introduced in Chapter 5.

Given the values of the parameters in the second part of the model the most likely values of the baseline rates, λ^t_{C}, were shown in Section 30.2 to be

$$\frac{d^t}{\sum_i y^t_i \theta_i}.$$

where θ_i is given by the second part of the model. When we divide time into clicks of duration h and redefine y^t_i to be 0 or 1 at-risk indicators, this expression becomes

$$\frac{d^t}{\sum_i h y^t_i \theta_i}.$$

In most clicks no failure occurs, $d^t = 0$, and the estimate of the rate is zero. In a click in which a failure occurs, $d^t = 1$, the estimated rate is

$$\frac{1}{h \sum_i y^t_i \theta_i},$$

which becomes very large as h becomes very small. However, the *cumulative* baseline rate increases at each click by the amounts $h\lambda^t_{\mathrm{C}}$, and the estimated values of these are either zero or

$$\frac{1}{\sum_i y^t_i \theta_i}$$

when a failure occurs. Thus the cumulative baseline rate is estimated by stepped curve with jumps at the observed failure times. This is called the Aalen–Breslow estimate and is illustrated in Fig. 30.2. The height of the jump at each failure time is now given by

$$1 \Big/ \sum_{\text{Risk set}} \theta$$

rather than by

$$1/(\text{Number of subjects at risk})$$

as in the simpler case discussed in Chapter 5. As noted there, examination of the cumulative rate plot allows us to assess the dependence of failure rate on time.

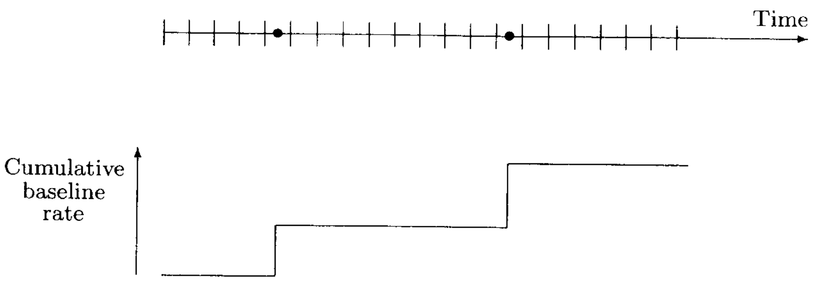

Fig. 30.2. The Aalen–Breslow estimate of the cumulative baseline rate.

Solutions to the exercises

30.1 When date is the time scale, membership of risk sets is determined by whether or not the subject was observed at the date of occurrence of the death. The risk sets corresponding to the two deaths are as follows:

Date of death	Subjects in risk set
3/10/68	*A, C, D, E, G (case)*
4/ 7/79	*A, B, C, D, E (case), F, H, I, J*

The risk set corresponding to the death of subject *G* contains fewer individuals since it occurred at a date earlier than some subjects had joined the cohort.

When age is the time scale, risk set membership is determined by whether the subject was observed at the age at which the death occurred. The risk sets are now as follows:

Age at death	Subjects in risk set
50.4	*A, C, D, E, F, G (case),I*
52.6	*A, C, D, E (case), F, I, J*

When time in study is the scale, the risk sets are as follows:

Time in study at death	Subjects in risk set
15.2 yrs	*A, B, C, D, E, G (case), H, I, J*
19.5 yrs	*A, C, D, E (case), H*

30.2 Since subject G is female and subject E is male, the risk set for the failure of G contains only female subjects and risk sets for the failure of E contains only males. When date is the time scale, the risk sets corresponding to the two deaths are as follows:

Date of death	Subjects in risk set
3/10/68	*A, D, G (case)*
4/ 7/79	*B, C, E (case), H, I*

When age is the time scale, the risk sets are

Age at death	Subjects in risk set
50.4	*A, D, G (case)*
52.6	*C, E (case), F, I*

When time in study is the scale, the risk sets are:

Time in study at death	Subjects in risk set
15.2 yrs	*A, D, G (case), J*
19.5 yrs	*C, E (case), H*