31
Time-varying explanatory variables

Cox's method provides a convenient way of controlling for time in the analysis of follow-up studies. In its simple form the method assumes that other explanatory variables do not change with time. In this chapter we show how the method can be extended to allow for this. We also discuss the closely related problem of analysis strategies when rates vary in relation to more than one time scale, and draw attention to some dangers and difficulties.

31.1 The model and the likelihood

We have seen that Cox's method amounts to dividing the multiplicative model for rates into two parts:

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

The first part refers to the baseline rates while the second part specifies how the rate ratio

$$\theta_i = \frac{\text{Rate for subject } i \text{ at time } t}{\text{Baseline rate at time } t}$$

is related to the explanatory variables A, B, etc.. On a log scale

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{\text{A} + \text{B} + \cdots}\,.$$

In the simple form of the method θ_i is assumed to be independent of time.

The extension of Cox's method with which we are now concerned allows the relationship between θ_i and the explanatory variables to vary with time. This would be necessary, for example, when studying levels of hazardous industrial exposures in occupational studies and when studying changing treatments in long term follow-up studies of chronic disease aetiology. Indeed *most* explanatory variables of interest to epidemiologists vary with time if follow-up is over a sufficiently long period.

Allowing the rate ratio part of the model to change over time involves

only a simple change to the contribution

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

from each risk set to the partial log likelihood. Since the model now predicts different values of θ at different times the contribution of each risk set must now be calculated using the values of θ current at the time of occurrence of the failure.

COMPUTATION

When it comes to computing the likelihood and finding the values of parameters which maximize it this simple change turns out to have major consequences, and computation times can increase by several orders of magnitude. To understand why the computation is so heavy it helps to look at the simpler version of Cox's method to see why this does *not* involve heavy computations. There are two reasons. First, for any particular set of values for the parameters, the value of θ only needs to be worked out once for each subject. Second, the value of $\sum\theta$ does not have to be calculated from scratch for each risk set because the equivalent term from the previous risk set can be updated by subtracting the values of θ for all subjects lost to follow-up in the intervening period and adding the contributions of those newly joining the cohort. Other terms needed in the computation of gradient and curvature of the log likelihood can be updated in a similar way.

When the model allows the rate ratios θ to change over time a subject who appears in several risk sets can have different values of θ in each. This means that not only must the values of θ be re-calculated for each risk set but $\sum\theta$ and other gradient and curvature terms must be calculated from scratch. The result is that the computing time rises dramatically.

Some reduction in computing time can be achieved by sampling the risk sets. The algebraic equivalence of the partial likelihood in Cox's method and the conditional likelihood for matched case-control studies means that analyzing a cohort study using Cox's method is the same as analyzing it as a case-control study in which each incident case is individually matched with a control set in which the controls are all other subjects under study at the moment of incidence. Since a case-control study which draws many controls for each case provides very little more information than one which draws only a few, we shall lose little by taking a random sample of controls drawn from each risk set rather than using the entire risk set. Sampling risk sets in this way creates what is called a *nested case-control study*. Such studies offer a number of practical advantages in addition to considerable computational savings and will be discussed further in Chapter 33.

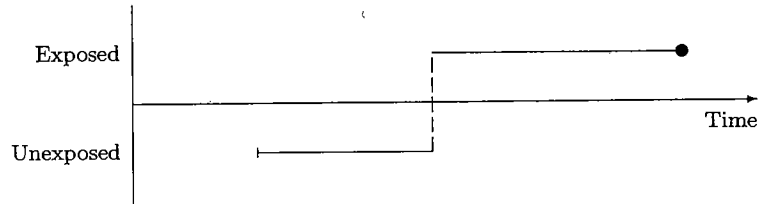

Fig. 31.1. Changing exposure group.

31.2 Changing exposure group

One simple but important way in which an explanatory variable can change with time arises when a subject can change from being unexposed to being exposed group (or vice versa) during the course of follow-up (see Fig. 31.1). This is most easily dealt with by splitting the follow-up for such subjects into an unexposed part and an exposed part, and treating the parts as distinct subjects. The data can then be analysed using the simple form of Cox's method in which the explanatory variables do not change with time. The validity of the analysis depends on a relatively strong assumption concerning the *reasons* for the change of exposure group, namely that transfer is unrelated to the subsequent probability of failure. If the transfer mechanism operates in a way that selects particularly high or low risk subjects then subsequent comparisons will be distorted. This is another example of selection bias. More formally, it is required that transfer must be independent of subsequent failure conditional upon the values of all other variables in the model. If transfer and failure are both strongly related to age (say) there will be an overall association between transfer time and outcome, but this will not bias estimates of other effects providing there is no relationship between transfer time and outcome *for subjects of the same age*, and providing the model takes proper account of the relationship between age and failure rate. Similar considerations apply when there are more than two categories of exposure or when the level of exposure varies continuously.

Exercise 31.1. Subjects enter a heart transplant programme as unexposed on joining a waiting list for a transplant, and switch to the exposed group on receiving the transplant. Do you think the assumptions discussed above are likely to be met in this case?

31.3 Time scales as explanatory variables

Another very common form of time-dependent explanatory variable is an additional time scale. For example, in a clinical study in which survival

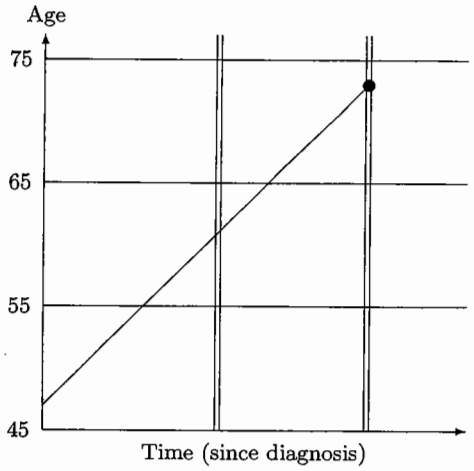

Fig. 31.2. Follow-up by age and time.

is analysed largely in relation to time since diagnosis, it will usually be necessary to control the comparison of different treatments for the age of the subjects receiving them. For short studies this can be achieved by including age at diagnosis, which is fixed for every subject. When follow-up is over many years it is better to include age itself, which varies with time. Fig. 31.2 illustrates follow-up of a subject in which observation time is classified by time since diagnosis and age. The risk sets are determined by the times of occurrence of failures. Two such times are illustrated in the figure by narrow vertical bands. One corresponds to the risk set for the failure of the subject shown while the other is an earlier failure. The subject shown contributes to both risk sets, but is of a different age on the two occasions.

One possible analysis would be to include time since diagnosis in the first part of the model, so that this is the time scale which is used to determine the risk sets, and to include age as a time varying explanatory variable in the second part of the model. This could be done either by dividing the age scale into 5- or 10-year bands and treating it as a categorical variable, as in

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{\text{Age} + \text{A} + \text{B} + \cdots},$$

or by treating age as a quantitative and fitting linear effects, and possibly quadratic effects too, as in

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{[\text{Age}] + [\text{Age-sq}] + \text{A} + \text{B} + \cdots}.$$

When the partial log likelihood is formed for either of these analyses each risk set contributes a term of the form $\log(\theta/\sum\theta)$ where the values of θ for the subjects in the risk set are determined by the relationship between $\log(\theta)$ and the parameters in the second part of the model. As an example of this computational process consider the model

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{\text{Age} + \text{A} + \text{B}}$$

where age has five levels, A has two levels and B has three levels. The parameters in the second part of the model are then Age(1), $\cdots$, Age(4), A(1), B(1) and B(2). Now consider a subject, at level 1 for A and level 2 for B, who appears as a survivor in the risk sets at two failure times, and suppose that this subject is in age band 3 at the time of the first failure, and in age band 4 at the time of the second failure.

Exercise 31.2. Write down an expression, in terms of the parameters, for the values of $\log(\theta)$ for this subject, in the two risk sets.

When there are two time scales a natural question to be considered is which should be included in the baseline rates part of the model and which should be included in the rate ratio part. The choice depends on the way that rates vary along each time scale. If this variation is to be modelled in the rate ratio part of the model then we must either divide the scale into broad bands or fit simple mathematical functions of time, such as linear or quadratic. The former strategy is adequate if the variation of rates is not too rapid, while the latter is only possible if the variation is regular enough to describe by simple mathematical functions. If variation is both rapid and irregular neither approach works very well and the variation should be modelled in the baseline rates. Thus if it is suspected that variation along one scale will be rapid and irregular this should be the scale whose effects are modelled by the baseline rates, and other scales should be treated as time varying explanatory variables. If variation is smooth along all scales it is better to use the scale with the strongest effects for the baseline rates.

Exercise 31.3. Discuss appropriate strategies for modelling the effects of age and calendar time on incidence of (a) a chronic degenerative disease, and (b) an infectious disease.

31.4 Dependencies between time scales ⋆

Different time scales are not truly different variables but the same variable measured from different origins. It is therefore impossible for a subject to advance one year on one scale without simultaneously advancing one year on other time scales. For example, we cannot pass through a year of calendar time without advancing a year in age — would that we could! This dependency between time scales can lead to difficulties when trying to interpret the estimated effects of changes on these time scales.

As an illustration we shall return to the example of age and time since diagnosis in a clinical follow-up study. Let us first consider the model

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{[\text{Age-at-diagnosis}] + \cdots},$$

in which the effect of time since diagnosis is the main time scale and is included in the first part of the model, while age at diagnosis is included as a linear effect in the second. The parameter [Age-at diagnosis] measures the change in the log rate per one year change in age, holding time since diagnosis constant at any arbitrary value. Fig. 31.3 shows two subjects who are diagnosed at ages 47 and 61 respectively; if we assume these subjects have the same values for any other explanatory variables the difference in log rate predicted by the model, at diagnosis, or at any value of time since diagnosis, is

$$(61 - 47) \times [\text{Age-at-diagnosis}] = 14 \times [\text{Age-at-diagnosis}].$$

Now consider the model

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{[\text{Age}] + \cdots}$$

in which age varies with time. The two subjects in Fig. 31.3 have a 14 year age difference at diagnosis, so this model predicts a difference in log rates between the two subjects of $14 \times$ [Age] at diagnosis. Because these two subjects have a 14 year age difference not only at diagnosis but at any time after diagnosis, the model also predicts a difference of $14 \times$ [Age] at any value of time since diagnosis. Thus both models predict a constant difference in log rate at any value of time since diagnosis. In the one case the prediction is $14 \times$ [Age-at-diagnosis], in the other the prediction is $14 \times$ [Age]. This is true for any pair of subjects; the models make identical predictions and cannot be differentiated, the [Age-at-diagnosis] parameter in the first model is making the same comparison as the [Age] parameter in the second.

There may well be scientific interest in discriminating between models in which the age at diagnosis determines prognosis, and models in which age itself is the determinant, but if we were to fit the model

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{[\text{Age}] + [\text{Age-at-diagnosis}] + \cdots},$$

in order to try and separate the linear effect of age controlled for time since diagnosis from the linear effect of age at diagnosis controlled for time since diagnosis, we would run into difficulties. When time since diagnosis and age are held constant, there can be no further variation in age at diagnosis so that the [Age-at-diagnosis] parameter cannot be estimated. Likewise,

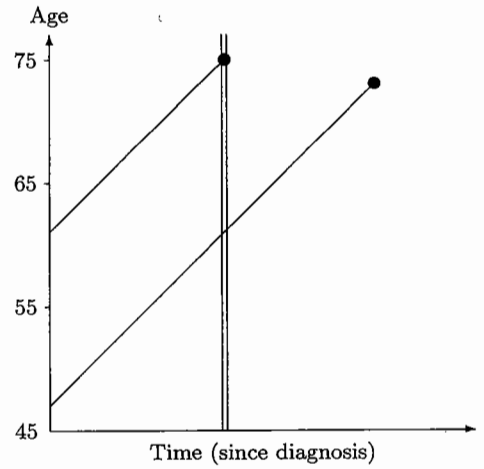

Fig. 31.3. Observation of two subjects.

time since diagnosis and age at diagnosis uniquely determine age so that the [Age] parameter cannot be estimated. Again the two subjects shown in Fig. 31.3 demonstrate the problem. The new model also predicts that the difference in log rates remains constant at any value of time since diagnosis but this difference is now equal to

$$14 \times [\text{Age}] + 14 \times [\text{Age-at-diagnosis}] = 14 \times ([\text{Age}] + [\text{Age-at-diagnosis}]),$$

where the parameters [Age] and [Age-at-diagnosis] now refer to the new model which contains both linear effects. Because any values for the two parameters which have the same sum, make the same predictions, the parameters cannot be estimated individually. They are said to be *non-identifiable* or *aliased.*

A computer program will usually warn the user when two parameters are non-identifiable and then omit one of them from the model. This is quite useful when the object is to control for age and age at diagnosis, but if the object is to disentangle their effects, what the computer program is saying is that we are attempting the impossible.

The non-identifiability of parameters for different time scales refers to their linear effects. When we come to fit models with non-linear terms, things are not so bad. Consider for example the predictions of the model

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{[\text{Age}] + [\text{Age-sq}] + \cdots}$$

for the two subjects shown in Fig. 31.3. At the time of diagnosis the model predicts a difference in log rates of

$$(61 - 47) \times [\text{Age}] + (61^2 - 47^2) \times [\text{Age-sq}] = 14 \times [\text{Age}] + 1512 \times [\text{Age-sq}].$$

However, 5 years after diagnosis, their ages are 66 and 52 and the model predicts a difference in log rates of

$$(66 - 52) \times [\text{Age}] + (66^2 - 52^2) \times [\text{Age-sq}] = 14 \times [\text{Age}] + 1652 \times [\text{Age-sq}].$$

In the model with non-linear effects, therefore, the difference between log rates for the two subjects does vary with time since diagnosis. The model

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{[\text{Age-at-diagnosis}] + [\text{Age-at-diagnosis-sq}] + \cdots}$$

predicts a difference in log rates of

$$(61 - 47) \times [\text{Age-at-diagnosis}] + (61^2 - 47^2) \times [\text{Age-at-diagnosis-sq}]$$

throughout the follow-up, and this is a different prediction than the one obtained from the model with age and age-squared. The linear parts of the two predictions are still the same and cannot be separately estimated, but the non-linear parts are different and can be.

Similarly, if we were to fit the model

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{\begin{array}{l}[\text{Age}] + [\text{Age-sq}] + [\text{Age-at-diagnosis}] + \\ [\text{Age-at-diagnosis-sq}] + \cdots\end{array}},$$

the parameters [Age] and [Age-at-diagnosis] are not identifiable while the parameters [Age-sq] and [Age-at-diagnosis-sq] can be estimated. The same is true for any other non-linear component of the relationships.

⋆ 31.5 Discrete time bands

In the above discussion the time variables are measured exactly; when the time scales are divided into discrete bands the position is slightly more complicated. To illustrate this we shall return to the two subjects of Fig. 31.3 and imagine a model in which age has been grouped into 5-year bands but time since diagnosis is still measured exactly. At the beginning of follow-up one subject is in the 45–49 band and the other is in the 60–64 band. However, after three years the former subject has moved into the 50–54 band while the latter remains in the 60–64 band. It will appear to a computer program that the age difference between the subjects has narrowed! As a result the program will not spot the underlying non-identifiability of models such as

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{\text{Age} + \text{Age-diag} + \cdots}$$

and fit them without complaint. However, the linear components of the relationships with age and age at diagnosis have only become estimable because of the inaccuracy introduced by banding and the resulting parameter estimates are uninterpretable.

31.6 Modelling vital rates ⋆

A familiar example of these problems arises in 'age-period-cohort' modelling of mortality and other vital rates, where the aim is to disentangle the dependence of rates upon age, calendar time (period), and date of birth (birth cohort). This comparison raises exactly the same problem as above and has provoked a lot of discussion in the epidemiological literature. Much of this has been based on the misconception that the problem is a shortcoming of current statistical methods and that its solution awaits only methodological advances. This is not the case. The difficulty is inescapable and arises from the fact that subjects cannot move in one time scale without an identical move in others.

Fig. 31.4 shows a table in which both both age and calendar period have been divided into 10-year bands. Tables of rates, classified in this way, are frequently available from official published sources, and allow effects of year of birth (*birth cohort* effects) to be estimated approximately. If we remember that observation of individual subjects is represented by diagonal lines in the age and calendar time Lexis diagram (illustrated by the arrow), it is clear that diagonal groupings of cells in the table correspond *approximately* to birth cohorts. The cell labelled 0 refers to subjects born around 1870, those labelled 1 to subjects born around 1880, and so on. Although this correspondence is only approximate, the new discrete codings for age period and cohort behave very much like the underlying continuous scales. In particular, they are linearly dependent. In our example,

$$\text{Cohort} = 3 + \text{Period} - \text{Age}.$$

This means that when two are fixed the third is also fixed and in models such as

$$\log(\text{Rate}) = \text{Corner} + [\text{Age}] + [\text{Period}] + [\text{Cohort}]$$

the parameters are unidentifiable, and it is impossible to disentangle the linear effects of all three variables.

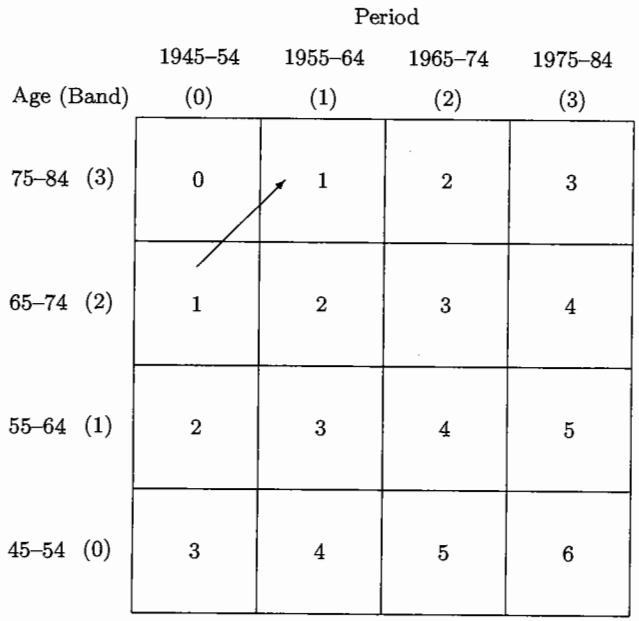

Fig. 31.4. Approximate birth cohorts.

Some investigators have returned to the raw data in order to allocate subjects to their true birth cohort. This avoids the approximation in Fig. 31.4 but leads to a serious fallacy. Fig. 31.5 shows how the exact birth cohorts move across the Lexis diagram. The cell labelled 0 refers to the 1860–69 birth cohort, those labelled 1 to the 1870–79 cohort, and so on. The discrete codings no longer behave like the underlying scales. For example, birth cohort 1 is observed in 3 cells; the transition from the first to the second involves a change of age band (from 65–74 to 75–84) without change in calendar period, while the transition from second to third corresponds to a move through calendar time without change in age! Looked at naively it would appear that, by grouping, we have created a natural experiment in which subjects can age instantaneously and travel in time without ageing. The fallacy lies in the fact that the regions are triangular and that regions shaped ▽ disproportionately represent ages towards the upper end of the 10-year band and dates towards the lower end of the period, while regions shaped △ disproportionately represent ages at the lower end of the band and periods at the upper end. Unfortunately, computer programs have no way of knowing this. They will believe that a miraculous natural experiment has been observed, and estimate separate linear effects for all

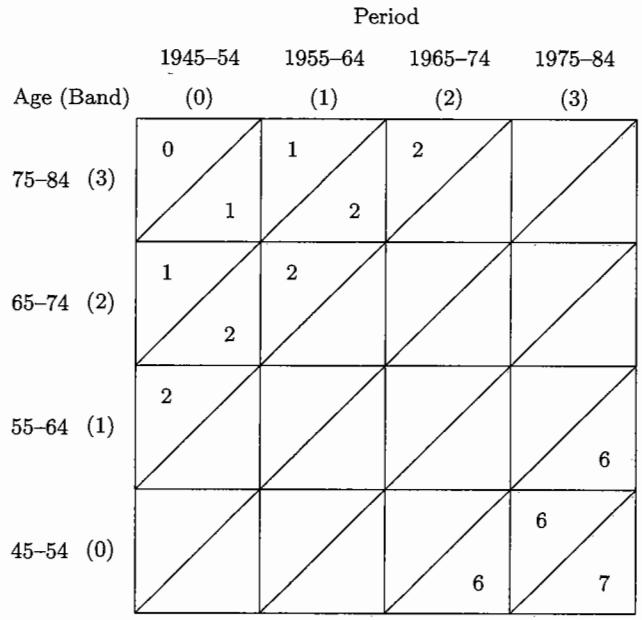

Fig. 31.5. Exact birth cohorts.

three scales without complaint. This uncritical behaviour of computer programs (which can't know better) has been hailed by some epidemiologists and statisticians (who should) as a 'solution' to the identifiability 'problem'. The reverse is the case; the computer solution is fallacious, being based entirely on grouping inaccuracies, and the resultant estimates are uninterpretable. It is worth pointing out that this pitfall is not confined to the age-period-cohort problem, but can be encountered whenever more than one time scale is involved in an analysis.

Solutions to the exercises

31.1 When a heart becomes available for transplantation and there is more than one patient eligible to receive it, there is potential selection bias. A controlled study would *randomize* such choices to exclude selection bias, but in an observational study it will always be difficult to know whether the recipient was selected because the clinician felt that this patient was most likely to benefit. Such selection would cause serious bias in a simple analysis. In theory this can be offset by including in the analysis any prognostic factors likely to have been used by the clinician making the decision, but in practice one can rarely be sure that all relevant factors

have been taken into account. We shall discuss this example in more detail in Chapter 32.

31.2 For the first risk set

$$\log(\theta) = \text{Age}(3) + \text{A}(1) + \text{B}(2).$$

For the second risk set

$$\log(\theta) = \text{Age}(4) + \text{A}(1) + \text{B}(2).$$

31.3 Incidence rates of chronic degenerative diseases such as ischaemic heart disease and most cancers rise steeply with age. In such diseases age may usually be thought of as a surrogate for the cumulative damage inflicted by a large number of influences throughout life. Such cumulative damage will be reflected in a *smooth* increase of rates with age so that simple linear or quadratic models for the age effect are usually satisfactory. Grouping age by 5 or 10 year bands will also work quite well. Age relationships for incidence of infectious diseases are usually more complicated. Increasing immunity with age will produce a smoothly decreasing curve, but where transmission of the infectious agent depends upon various social influences such as schooling, employment, sexual activity etc., these may give rise to rather irregular age curves. Simple mathematical functions for age-incidence curves are therefore less likely to be useful. Grouping may also be difficult because of abrupt changes in incidence due to age related changes in social behaviour.