Chapter 6

Regression models and subject heterogeneity

6.1 Summary

We consider several models that describe survival in the presence of observable covariates, these covariates measuring subject heterogeneity. The most general situation can be described by a model with a parameter of high, possibly unbounded, dimension. Proportional hazards models, partially proportional hazards models (O'Quigley and Stare 2002), stratified models or models with frailties or random coefficients all arise as special cases of this model (O'Quigley and Xu 2000). One useful parameterization (O'Quigley and Pessione 1991, O'Quigley and Prentice 1991) can be described as a non proportional hazards model with intercept. Changepoint models are a particular form of a non proportional hazards model with intercept (O'Quigley and Natarajan 2004). Any model can be viewed as a special case of the general model, lying somewhere on a conceptual scale between this general model and the most parametric extreme, which would be the simple exponential model. Models can be placed on this scale according to the extent of model constraints and, for example, a random effects model would lie strictly between a stratified model and the simple exponential model. Relative risk models used in epidemiology come under these headings. For relative risk models the time component is usually taken to be age and great generalization, e.g., period or cohort analysis is readily accomplished. Time-dependent covariates, $Z(t)$, in combination with the at-risk indicator, $Y(t)$, can be used to describe states. Multistate models in which subjects can move in and out of different states, or

into an absorbing state such as death, can then be analyzed using the same methodology.

6.2 Motivation

The presence of subject heterogeneity, summarized by risk factors Z, known or suspected of being related to $S(t)$, is our central concern. The previous chapter dealt with the issue of marginal survival, i.e., survival ignoring any indicator of heterogeneity and which treats the data in hand as though the observations came from a single population. In Figure 6.1 there are two groups. This can be described by two distinct Kaplan-Meier curves or, possibly, two independently calculated fitted parametric curves. If, however, the curves are related, then each estimate provides information not only about its own population curve but also about the other group's population curve. The curve estimates would not be independent. Exploiting such dependence can lead to considerable gains in our estimating power. The agreement between an approach modeling dependence and one ignoring it can be more or less strong and, in Figure 6.1, agreement is good apart from observations beyond 150 months where a proportional hazards assumption may not hold very well. Returning to the simplest case, we can imagine a compartmental model describing the occurrence of

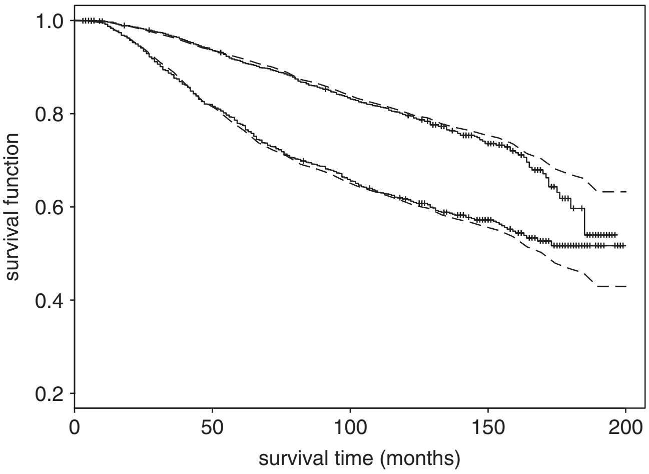

Figure 6.1: Kaplan-Meier survival curves and PH model curves for two groups defined by a binary covariate. Dashed lines represent PH estimates.

deaths independently of group status in which all individuals are assumed to have the same hazard rates. As pointed out in the previous chapter, the main interest then is in the survival function $S(t)$ when the Z are either unobservable or being ignored. Here we study the conditional survival function given the covariates Z and we write this as $S(t|Z)$. In the more complex situations (multicompartment models, time-dependent Z) it may be difficult, or even impossible, to given an interpretation to $S(t)$ as an average over conditional distributions, but the idea of conditioning is still central although we may not take it beyond that of the probability of a change of state conditional upon the current state as well as the relevant covariate history which led to being in that state.

The goal here is to consider models with varying degrees of flexibility applied to the summary of n subjects each with an associated covariate vector Z of dimension p. The most flexible models will be able to fully describe any data at hand but, as a price for their flexibility, little reduction in dimension from the $n \times p$ data matrix we begin with. Such models will have small bias in prediction compared with large sampling errors. The most rigid models can allow for striking reductions in dimension. Their consequent impact on prediction will be associated with much smaller sampling errors. However, as a price for such gains, the biases in prediction can be large. The models we finally work with will lie between these two extremes. Their choice then depends on an artful balance between the two conflicting characteristics.

6.3 General or nonproportional hazards model

In the most straightforward cases we can express the conditional dependence of survival upon fixed covariates in terms of the hazard function. A general expression for the hazard function given the value of the covariate Z is given by:

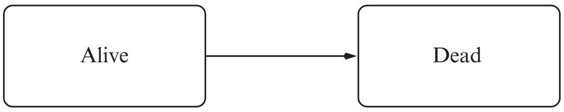

Figure 6.2: A simple alive/dead transition model. At time t the only information being used is whether the subject is dead or alive. Covariate information (eg. group status) is not used.

$$\lambda(t|Z) = \lambda_0(t)\exp\{\beta(t)Z\}, \tag{6.1}$$

where $\lambda(t|\cdot)$ is the conditional hazard function, $\lambda_0(t)$ the baseline hazard corresponding to $Z = 0$, and $\beta(t)$ a time-varying regression effect. Whenever Z has dimension greater than one we view $\beta(t)Z$ as an inner product in which $\beta(t)$ has the same dimension as Z so that $\beta(t)Z = \beta_1(t)Z_1+, \cdots, +\beta_p(t)Z_p$.

Recalling the discussion of Chapter 5, we are only interested in situations where observations on Z can be made in the course of any study. In Equation 6.1 Z is not allowed to depend upon time. If we also disallow the possibility of continuous covariates, which, in practice, we can approximate as accurately as we wish via high dimensional Z together with $\beta(t)$ of the same dimension, we see that model (6.1) is completely general and, as such, not really a model. It is instead a representation, or re-expression, of a very general reality, an expression that is convenient and which provides a framework to understanding many of the models described in this chapter. At the cost of losing the interpretation of a hazard function, we can immediately generalize (6.1) to

$$\lambda(t|Z) = \lambda_0(t)\exp\{\beta(t)Z(t)\}. \tag{6.2}$$

As long as we do not view $Z(t)$ as random, i.e., the whole time path of $Z(t)$ is known at $t = 0$, then a hazard function interpretation for $\lambda(t|Z)$ is maintained. Otherwise we lose the hazard function interpretation, since this requires knowledge of the whole function at the origin $t = 0$, i.e., the function is a deterministic and not a random one. In some ways this loss is of importance in that the equivalence of the hazard function, the survival function, and the density function means that we can easily move from one to another. However, when $Z(t)$ is random, we can reason in terms of intensity functions and compartmental models, a structure that enables us to deal with a wide variety of applied problems. The parameter $\beta(t)$ is of infinite dimension and therefore the model would not be useful without some restrictions upon $\beta(t)$.

6.4 Proportional hazards model

Corresponding to the truth or reality under scrutiny, we can view Equation (6.2) as being an extreme point on a large scale which calibrates model complexity. The opposite extreme point on this scale might have been the simple exponential model, although we will start with a restriction that is less extreme, specifically the proportional hazards model in which $\beta(t) = \beta$ so that;

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta Z(t)\}. \tag{6.3}$$

Putting restrictions on $\beta(t)$ can be done in many ways, and the whole art of statistical modeling, not only for survival data, is in the search for useful restrictions upon the parameterization of the problem in hand. Our interpretation of the word "useful" depends very much on the given particular context.

Just where different models find themselves on the infinite scale between Equation 6.3 and Equation 6.2 and how they can be ordered is a very important concept we need master if we are to be successful at the modeling process, a process which amounts to feeling our way up this scale (relaxing constraints) or down this scale (adding constraints), guided by the various techniques at our disposal. From the outset it is important to understand that the goal is not one of establishing some unknown hidden truth. We already have this, expressed via the model described in Equation (6.1). The goal is to find a much smaller, more restrictive model, which, for practical purposes is close enough or which is good enough to address those questions that we have in mind; for example, deciding whether or not there is an effect of treatment on survival once we have accounted for known prognostic factors which may not be equally distributed across the groups we are comparing. For such purposes, no model to date has seen more use than the Cox regression model.

6.5 The Cox regression model

In tackling the problem of subject heterogeneity, Cox's (1972) proportional hazards regression model has enjoyed outstanding success, a success, it could be claimed, matching that of classic multilinear regression itself. The model has given rise to considerable theoretical work and continues to provoke methodological advances. Research and development into the model and the model's offspring have become so extensive that we cannot here hope to cover the whole field, even at the time of writing. We aim nonetheless to highlight what seem to be the essential ideas and we begin with a recollection of the seminal paper of D.R. Cox, presented at a meeting of the Royal Statistical Society in London, England, March 8, 1972.

Regression models and life tables (D.R. Cox 1972)

After summarizing earlier work on the life table (Kaplan and Meier 1958, Chiang 1968), Professor Cox introduced his, now famous, model postulating a simplified form for the relationship between the hazard function $\lambda(t)$, at time t and the value of an associated fixed covariate Z. As its name suggests, the proportional hazards model assumes that the hazard functions among subjects with different covariates are proportional to one another. The hazard function can then be written:

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta Z\}, \tag{6.4}$$

where $\lambda_0(t)$ is a fixed "baseline" hazard function, and β is a relative risk parameter to be estimated. Whenever $Z = 0$ has a concrete interpretation (which we can always obtain by recoding) then so does the baseline hazard $\lambda_0(t)$ since, in this case, $\lambda(t|Z = 0) = \lambda_0(t)$. As mentioned just above, when Z is a vector of covariates, then the model is the same, although with the scalar product βZ interpreted as an inner product. It is common to replace the expression βZ by $\beta' Z$ where β and Z are $p \times 1$ vectors, and $a'b$ denotes the inner product of vectors a and b. Usually, though, we will not distinguish notationally between the scalar and the vector inner product since the former is just a special case of the latter. We write them both as βZ. Again we can interpret $\lambda_0(t)$ as being the hazard corresponding to the group for which the vector Z is identically zero.

The model is described as a multiplicative model, i.e., a model in which factors related to the survival time have a multiplicative effect on the hazard function. An illustration in which two binary variables are used to summarize the effects of four groups is shown in Figure 6.3. As pointed out by Cox, the function (βZ) can be replaced by any function of β and Z, the positivity of $\exp(\cdot)$ guaranteeing that, for any hazard function $\lambda_0(t)$, and any Z, we can always maintain a hazard function interpretation for $\lambda(t|Z)$. Indeed it is not necessary to restrict ourselves to $\exp(\cdot)$, and we may wish to work with other functions $R(\cdot)$, although care is required to ensure that $R(\cdot)$ remains positive over the range of values of β and Z of interest. Figure 6.3 represents the case of two binary covariables indicating four distinct groups (in the figure we take the logarithm of $\lambda(t)$) and the important thing to observe is that the distance between any two groups on this particular scale, i.e., in terms of the log-hazards, does not change through time. In view of the relation between the hazard function and the survival function,

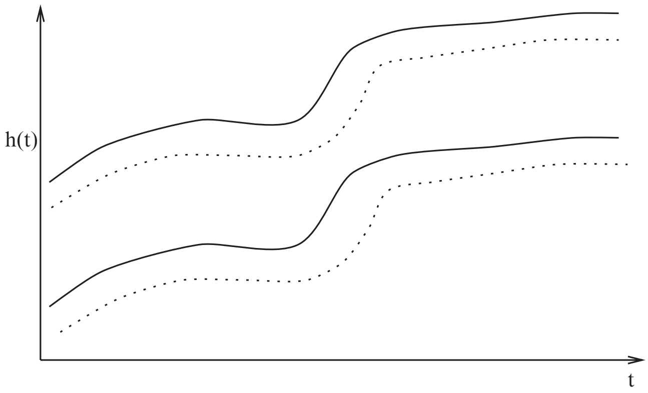

Figure 6.3: Proportional hazards with two binary covariates indicating 4 groups. Log-hazard rate written as $h(t) = \log\lambda(t)$.

there is an equivalent form of Equation 6.4 in terms of the survival function. Defining $S_0(t)$ to be the baseline survival function; that is, the survival function corresponding to $S(t|Z = 0)$, then, for scalar or vector Z, we have that,

$$S(t|Z) = \{S_0(t)\}^{\exp(\beta Z)}. \tag{6.5}$$

When the covariate is a single binary variable indicating, for example, treatment groups, the model simply says that the survival function of one group is a power transformation of the other, thereby making an important connection to the class of Lehmann alternatives (Lehmann 1953).

Cox took the view that "parametrization of the dependence on Z is required so that our conclusions about that dependence are expressed concisely," adding that any choice "needs examination in the light of the data." "So far as secondary features of the system are concerned ... it is sensible to make a minimum of assumptions." This view led to focusing on inference that allowed $\lambda_0(t)$ to remain arbitrary. The resulting procedures are nonparametric with respect to t in that inference is invariant to any increasing monotonic transformation of t, but parametric in as much as concerns Z. For this reason the model is often referred to as Cox's semi-parametric model. Let's keep in mind, however, that it is the adopted inferential procedures that are semi-parametric rather than the model itself. Although, of course, use of the term $\lambda_0(t)$ in the model, in which $\lambda_0(t)$ is not specified, implies use of procedures that will work for all allowable functions $\lambda_0(t)$.

Having recalled to the reader how inference could be carried out following some added assumptions on $\lambda_0(t)$, the most common assumptions being that $\lambda_0(t)$ is constant, that $\lambda_0(t)$ is a piecewise constant function, or that $\lambda_0(t)$ is equal to t^γ for some γ, Cox presented his innovatory likelihood expression for inference, an expression that subsequently became known as a partial likelihood (Cox 1975). We look more closely at these inferential questions in later chapters. First note that the quantity $\lambda_0(t)$ does not appear in the expression for partial likelihood given by

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta Z_i)}{\sum_{j=1}^{n} Y_j(X_i) \exp(\beta Z_j)} \right\}^{\delta_i} , \tag{6.6}$$

and, in consequence, $\lambda_0(t)$ can remain arbitrary. Secondly, note that each term in the product is the conditional probability that at time X_i of an observed failure, it is precisely individual i who is selected to fail, given all the individuals at risk and given that one failure would occur. Taking the logarithm in Equation 6.6 and its derivative with respect to β, we obtain the score function which, upon setting equal to zero, can generally be solved without difficulty using the Newton-Raphson method, to obtain the maximum partial likelihood estimate $\hat{\beta}$ of β. We will discuss more deeply the function $U(\beta)$ under the various approaches to inference. We can see already that it has the same form as that encountered in the standard linear regression situation where the observations are contrasted to some kind of weighted mean. The exact nature of this mean is described later. Also, even though the expression

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left\{ Z_i - \frac{\sum_{j=1}^{n} Y_j(X_i) Z_j \exp(\beta Z_j)}{\sum_{j=1}^{n} Y_j(X_i) \exp(\beta Z_j)} \right\} \tag{6.7}$$

looks slightly involved, we might hope that the discrepancies between the Z_i and the weighted mean, clearly some kind of residual, would be uncorrelated, at least for large samples, since the Z_i themselves are uncorrelated.

All of this turns out to be so and makes it relatively easy to carry out appropriate inference. The simplest and most common approach to inference is to treat $\hat{\beta}$ as asymptotically normally distributed with mean β and large sample variance $I(\hat{\beta})^{-1}$, where $I(\beta)$, called the information in view of the analogy with classical likelihood, is minus the second derivative of $L(\beta)$ with respect to β, i.e., letting

$$I_i(\beta)=\frac{\sum_{j=1}^n Y_j(X_i)Z_j^2\exp(\beta Z_j)}{\sum_{j=1}^n Y_j(X_i)\exp(\beta Z_j)}-\left\{\frac{\sum_{j=1}^n Y_j(X_i)Z_j\exp(\beta Z_j)}{\sum_{j=1}^n Y_j(X_i)\exp(\beta Z_j)}\right\}^2, \quad (6.8)$$

then $I(\beta) = \sum_{i=1}^n \delta_i I_i(\beta)$. Inferences can also be based on likelihood ratio methods. A third possibility, which is sometimes convenient, is to base tests on the score $U(\beta)$, which in large samples can be considered to be normally distributed with mean zero and variance $I(\beta)$. Multivariate extensions are completely natural, with the score being a vector and I an information matrix.

Early applications of the model

The first success of the model was in its use for the two-sample problem, i.e., testing the null hypothesis of no difference in the underlying true survival curves for two groups. In this case Cox (1972) showed that the test statistic $U(0)/\sqrt{I(0)}$ is formally identical to a test, later known under the heading of the log-rank test, obtained by setting up at each failure point a 2 × 2 contingency table, group against failed/survived, and combining the many 2 × 2 tables. As in a standard analysis of a single such contingency table we use the marginal frequencies to obtain estimates of expected rates under the null hypothesis of no effect. Assuming, as we usually do here, no ties we can obtain a table such as described in Table 6.1 in which, at time $t = X_i$ the observed failure occurs in group A and there are $n_A(t)$ and $n_B(t)$ individuals at risk in the respective groups.

The observed rates and the expected rates are simply summed across the distinct failure points, each of which gives rise to its own contingency table where the margins are obtained from the available risk sets at that time. From the above, if $Z_i = 1$ when subject i is in group A and zero otherwise, then elementary calculation gives that,

$$U(0) = \sum_{i=1}^n \delta_i\{Z_i - \pi(X_i)\}\,, \quad I(0) = \sum_{i=1}^n \delta_i \pi(X_i)\{1 - \pi(X_i)\}$$

Time point $t = X_i$	Group A	Group B	Totals
Number of failures	1	0	1
Number not failing	$n_A(t) - 1$	$n_B(t)$	$n_A(t) + n_B(t) - 1$
Total at risk	$n_A(t)$	$n_B(t)$	$n_A(t) + n_B(t)$

Table 6.1: 2×2 table at failure point $t = X_i$ for group A and group B.

where $\pi(t) = n_A(t)/\{n_A(t)+n_B(t)\}$. The statistic U then contrasts the observations with their expectations under the null hypothesis of no effect. This expectation is simply the probability of choosing, from the subjects at risk, a subject from group A. The variance expression is the well-known expression for a Bernoulli variable. Readers interested in a deeper insight into this test should also consult (Cochran 1954, Mantel and Haenzel 1959, Mantel 1963, Peto and Peto 1972). As pointed out by Cox, "whereas the test in the contingency table situation is, at least in principle, exact, the test here is only asymptotic ..."

However, the real advantage of Cox's approach was that while contributing significantly toward a deeper understanding of the log-rank and related tests, it opened up the way for more involved situations; additional covariates, continuous covariates, random effects and, perhaps surprisingly, in view of the attribute "proportional hazards," a way to tackle problems involving time varying effects or time dependent covariates. Cox illustrated his model via an application to the now famous Freireich data (Freireich et al. 1963) describing a clinical trial in leukemia in which a new treatment was compared to a placebo. Treating the two groups independently and estimating either survivorship function using a Kaplan-Meier curve gave good agreement with the survivorship estimates derived from the Cox model. Such a result can also, of course, be anticipated by taking a $\log(-\log)$ transform of the Kaplan-Meier estimates and noting that they relate to one another via a simple shift. This shift exhibits only the weakest, if any, dependence on time itself.

Multivariate applications

Recovering the usual two-group log rank statistic as a special case of a test based on model (6.4) is reassuring. In fact, exactly the same approach extends to the several group comparison (Breslow 1972). More importantly, model (6.4) provides the framework for considering the multivariate problem from its many angles; global comparisons of course but also more involved conditional comparisons in which certain effects are controlled for while others are tested. We look at this in more detail below under the heading "Modeling multivariate problems." The partially proportional hazards model (in particular the stratified model) were to appear later to Cox's original work of 1972 and provide great flexibility in addressing regression problems in a multivariate context.

Discussion of Professor Cox's paper

Professor Cox's paper represented an important step forward in dealing with survival problems for heterogeneous populations and a nonnegligible subset of a whole generation of academic biostatisticians has spent over a quarter of a century, keeping up and clarifying the many ideas originally outlined in Cox's 1972 paper. The discussion continues but, already, back in 1972 a group drawn from among the most eminent statisticians of the time, made a collective contribution to the new developments in a discussion that turned out to be almost as significant as the paper itself.

The issue which, arguably, gave rise to the most fertile exchanges concerned the partial likelihood, not yet named as such and referred to by Cox as a conditional likelihood. Kalbfleisch and Prentice took issue with Cox's naming of the likelihood used for inference as a "conditional" likelihood. They pointed out that the likelihood expression is not obtainable as a quantity proportional to a probability after having conditioned on some event. Conditioning was indeed taking place in the construction of the likelihood expression but in a sequential manner, a dynamic updating whose inferential home would later be seen to lie more naturally within the context of stochastic processes, indexed by time, rather than regular likelihoods, whether marginal or conditional.

The years following this discussion gave rise to a number of papers investigating the nature of the "conditional" likelihood proposed in Cox's original paper. Given the striking success of the model, together with the suggested likelihood expression, in reproducing and taking further a wide range of statistics then in use, most researchers agreed that Cox's proposal was correct. They remained uncertain, though, as to how to justify the likelihood itself. This thinking culminated in several major contributions; those of Cox (1975), Prentice and Kalbfleisch (1975), Aalen (1979) and Andersen and Gill (1982), firmly establishing the likelihood expression of Cox. In our later chapter on inference we discuss some of the issues raised in those contributions. It turned out that Cox was correct, not just on the appropriateness of his proposed likelihood expression but also in describing it as a "conditional" likelihood, this description being the source of all the debate.

Not unlike other major scientific thinkers of the twentieth century, Cox showed quite remarkable insight and although his likelihood derivation may not have been conditional, in the sense of taking as observed some single statistic upon which we condition before proceeding, his likelihood is not only very much a conditional one but also it conditions in just the right way. Not in the most straightforward sense whereby all the conditioning is done in one go, but in the sense of sequentially conditioning through time. Cox's "conditional" likelihood is now called a "partial" likelihood although, as an inferential tool in its own right, i.e., as a tool for inference independent of the choice of any particular model the partial likelihood is not as useful a concept as believed by many. We return to this in the chapter on inference.

Professor Downton of the University of Birmingham and Professor Peto of the University of Oxford pointed out the connection to rank test procedures. Although the formulation of Cox allowed the user to investigate more complex structures, many existing set-ups, framed in terms of tests based on the ranks, could be obtained directly from the use of the Cox likelihood. The simplest example was the sign test for the median. Using permutation arguments, other tests of interest in the multivariate setting could be obtained, in particular tests analogous to the Friedman test and the Kruskal-Wallis test. Richard Peto referred to some of his own work with Julian Peto. Their work demonstrated the asymptotic efficiency of the log-rank test and that, for the two-group problem and for Lehmann alternatives, this test was locally most powerful. Since the log-rank test coincides with a score test based on Cox's likelihood, Peto argued that Cox's method necessarily inherits the same properties.

Professor Bartholomew of the University of Kent considered a lognormal model in current use and postulated its extension to the regression situation by writing down the likelihood. Such an analysis, being fully parametric, represents an alternative approach since the structure is not nested in a proportional hazards one. Bartholomew made an insightful observation that allowing for some dependence of the explanatory variable Z on t can enable the lognormal model and a proportional hazards model to better approximate each another. This is indeed true and allows for a whole development of a class of non proportional hazards models where Z is a function of time and within which the proportional hazards model arises as a special case.

Professors Oakes and Breslow discussed the equivalence between a saturated piecewise exponential model and the proportional hazards model. By a saturated piecewise exponential model we mean one allowing for constant hazard rates between adjacent failures. The model is data dependent in that it does not specify in advance time regions of constant hazard but will allow these to be determined by the observed failures. From an inferential standpoint, in particular making use of likelihood theory, we may expect to run into some difficulties. This is because the number of parameters of the model (number of constant hazard rates) increases at the same rate as the effective sample size (number of observed failure times). However, the approach does nonetheless work, although justification requires the use of techniques other than standard likelihood. A simple estimate of the hazard rate, the cumulative hazard rate, and the survivorship function are then available. When $\beta = 0$ the estimate of the cumulative hazard rate coincides with that of Nelson (1969).

Professor Lindley of University College London writes down the full likelihood which involves $\lambda_0(t)$ and points out that, since terms involving $\lambda_0(t)$ do not factor out we cannot justify Cox's conditional likelihood. If we take $\lambda_0(t)$ as an unknown nuisance parameter having some prior distribution, then we can integrate the full likelihood with respect to this in order to obtain a marginal likelihood (this would be different to the marginal likelihood of ranks studied later by Kalbfleisch and Prentice 1973). Lindley argues that the impact of censoring is greater for the Cox likelihood than for this likelihood which is then to be preferred. The author of this text confesses to not fully understanding Lindley's argument and there is some slight confusion there since, either due to a typo or to a subtlety that escapes me, Lindley calls the Cox likelihood a "marginal likelihood" and what I am referring to as a marginal likelihood, an "integrated likelihood." We do, of course, integrate a full likelihood to obtain a marginal likelihood, but it seems as though Professor Lindley was making other, finer, distinctions which are best understood by those in the Bayesian school. His concern on the impact of censoring is echoed by Mr. P. Glassborow of British Rail underlining the strength behind the independent censoring assumption, an assumption which would not be reasonable in many practical cases.

Professor Zelen, a pioneer in the area of regression analysis of survival data, pointed out important relationships in tests of regression effect in the proportional hazards model and tests of homogeneity of the odds ratio in the study of several contingency tables. Dr. John Gart of the National Cancer Institute also underlined parallels between contingency table analysis and Cox regression. These ideas were to be developed extensively in later papers by Ross Prentice and Norman

Breslow in which the focus switched from classical survival analysis to studies in epidemiology. The connection to epidemiological applications was already alluded to in the discussion of the Cox paper by Drs. Meshalkin and Kagan of the World Health Organization. Finally, algorithms for carrying out an analysis based on the Cox model became quickly available thanks to two further important contributions to the discussion of Cox's paper. Richard Peto obtained accurate approximations to the likelihood in the presence of ties, obviating the need for computationally intensive permutation algorithms, and Susannah Howard showed how to program efficiently by exploiting the nested property of the risk sets in reversed time.

Historical background to Cox's paper

Alternative hypotheses to a null which assumes that two probabilities are equal, such as in Equation (6.5), taking the form of a simple power transformation, have a long history in statistical modeling. Such alternatives which, in the special case where the probabilities in question are survival functions, are known as Lehmann alternatives (Lehmann 1953). Lehmann alternatives are natural in that, under the restriction that the power term is positive, always achievable by reparameterizing the power term to be of an exponential form; then, whatever the actual parameter estimates, the resulting probability estimates satisfy the laws of probability. In particular, they remain in the interval (0,1). Linear expressions for probabilities are less natural although, at least prior to the discovery of the logistic and Cox models, possibly more familiar. Feigl and Zelen (1965) postulated a linear regression for the location parameter, λ_0, of an exponential law. In this case the location parameter and the (constant) hazard coincide so that the model could be written;

$$\lambda(t|Z) = \lambda_0 \exp\{\beta Z\}. \tag{6.9}$$

In Feigl and Zelen their model was not written exactly this way, expressed as $\lambda = \alpha + \beta Z$. However, since λ is constant, the two expressions are equivalent and highlight the link to Cox's more general formulation. Feigl and Zelen only considered the case of uncensored data. Zippin and Armitage (1966) used a modeling approach, essentially the same as that of Feigl and Zelen, although allowing for the possibility of censoring. This was achieved by an assumption of independence between the censoring mechanism and the failure mechanism

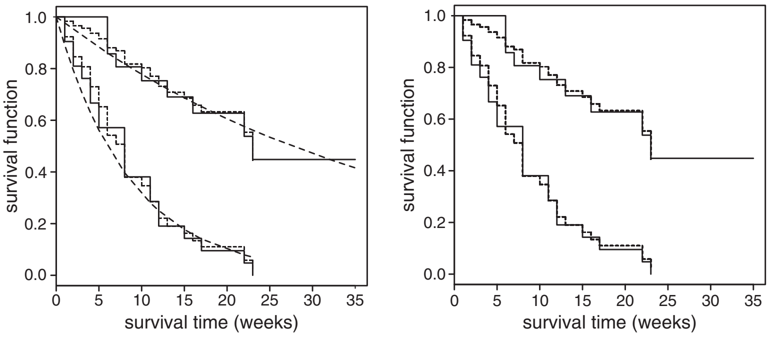

Figure 6.4: Kaplan-Meier curves and model based curves for Freireich data. Dashed lines represent model based estimates; exponential model (left), Cox model (right).

enabling an expression for the full likelihood to be obtained. Further discussion on these ideas can be found in Myers, Hankey and Mantel (1973) and Brown (1975). The estimates of the survival function for the different groups in the Freireich study, based on a simple exponential model or a Cox model, are shown in Figure 6.4. For these data the level of agreement between the two approaches appears to be high. This early work on the exponential model certainly helped anticipate the more general development of Cox and, for many more straightforward comparisons, such as the one illustrated by the Freireich data, it is perhaps unfortunate that the exponential model has been relegated to a historical role alone and is rarely, if ever, used in current practical analysis of similar data.

6.6 Modeling multivariate problems

The strength of the Cox model lies in its ability to describe and characterize involved multivariate situations. Crucial issues concern the adequacy of fit of the model, how to make predictions based on the model, and how strong is the model's predictive capability. These are considered in detail later. Here, in the following sections and in the chapter on inference we consider how the model can be used as a tool to formulate questions of interest to us in the multivariate setting. The simplest case is that of a single binary covariate Z taking the values zero and one. The zero might indicate a group of patients undergoing a standard therapy, whereas the group for which $Z = 1$ could be undergoing some experimental therapy. Model 6.4 then indicates the hazard rate for the standard group to be $\lambda_0(t)$ and for the experimental group to be $\lambda_0(t)\exp(\beta)$. Testing whether or not the new therapy has any effect on survival translates as testing the hypothesis $H_0 : \beta = 0$. If β is less than zero then the hazard rate for the experimental therapy is less than that for the standard therapy at all times and is such that the arithmetic difference between the respective logarithms of the hazards is of magnitude β. Suppose the problem is slightly more complex and we have two new experimental therapies. We can write;

$$\lambda(t|Z) = \lambda_0(t)\exp\{\beta_1 Z_1 + \beta_2 Z_2\}$$

and obtain Table 6.2. As we shall see the two covariate problem is very much more complex than the case of a single covariate. Not only do we need to consider the effect of each individual treatment on the hazard rate for the standard therapy but we also need to consider the effect of each treatment in the presence or absence of the other as well as the combined effect of both treatments together. The particular model form in which we express any relationships will typically imply assumptions on those relationships and an important task is to bring under scrutiny (goodness of fit) the soundess of any assumptions.

It is also worth noting that if we are to assume that a two-dimensional covariate proportional hazards model hold exactly, then, integrating over one of the covariates to obtain a one dimensional model will not result (apart from in very particular circumstances) in a lower-dimensional proportional hazards model. The lower dimensional model would be in a much more involved non proportional hazards form. This observation also holds when adding a covariate to a one-dimensional proportional hazards model, a finding that compels us, in realistic modeling situations, to only ever consider the model as an approximation.

By extension the case of several covariates becomes rapidly very complicated. If, informally, we were to define complexity as *the number*

Treatment group	Z_1	Z_2	Log of group effect
Standard therapy	0	0	0
Experimental therapy 1	1	0	β_1
Experimental therapy 2	0	1	β_2

Table 6.2: Effects for two treatment groups

of things you have to worry about, then we could, even more informally, state an important theorem.

Theorem 6.1 (Theorem of complexity) *The complexity of any problem grows exponentially with the number of covariates in the equation.*

Obviously such a theorem cannot hold in any precise mathematical sense without the need to add conditions and restrictions such that its simple take-home message would be lost. For instance, if each added covariate was a simple constant multiple of the previous one, then there would really be no added complexity. But, in some broad sense, the theorem does hold and to convince ourselves of this we can return to the case of two covariates. Simple combinatorial arguments show that the number of possible hypotheses of potential interest is increasing exponentially. But it is more complex than that. Suppose we test the hypothesis $H_0 : \beta_1 = \beta_2 = 0$. This translates the clinical null hypothesis: neither of the experimental therapies impacts survival against the alternative, $H_1 : \exists\, \beta_i \neq 0,\ i = 1, 2$. This is almost, yet not exactly, the same as simply regrouping the two experimental treatments together and reformulating the problem in terms of a single binary variable.

Next we might consider testing the null hypothesis $H_0 : \beta_1 = 0$ against the alternative hypothesis $H_1 : \beta_1 \neq 0$. Such a test focuses only on the first experimental treatment, but does not, as we might at first imagine, lump together both the second experimental treatment and the standard treatment. This test makes no statement about β_2 and so this could indeed take the value zero (in which case the standard and the second experimental therapy are taken to be the same) or any other value in which case, detecting a nonzero value for β_1 translates as saying that this therapy has an effect different to the standard regardless of the effect of the second experimental therapy. Clearly this is different from lumping together the second experimental therapy with the standard and testing the two together against the first experimental therapy. In such a case, should the effect of the first experimental therapy lie somewhere between that of the standard and the second, then, plausibly, we might fail to detect a nonzero β_1 even though there exist real differences between the standard and the first therapy.

All of this discussion can be repeated, writing β_1 in the place of β_2. Already, we can see that there are many angles from which to consider an equation such as the above. These angles, or ways of expressing the scientific question, will impact the way of setting up the statistical hypotheses. In turn, these impact our inferences.

Another example would be testing the above null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ against an alternative $H_1 : 0 < \beta_1 < \beta_2$ instead of that initially considered (i.e., $H_1 : \exists\, \beta_i \neq 0,\ i = 1, 2$). The tests, and their power properties, would not typically be the same. We might consider recoding the problem, as in Equation 6.10, so that testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ corresponds to testing for an effect in either group. Given this effect we can test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$ which will answer the question as to whether, given that their exists a treatment effect, it is the same for both of the experimental treatments:

$$\begin{aligned} \lambda(t|Z) &= \lambda_0(t) \exp\{\beta_1 Z_1 + (\beta_1 + \beta_2) Z_2\} \\ &= \lambda_0(t) \exp\{\beta_1 (Z_1 + Z_2) + \beta_2 Z_2\}. \qquad (6.10) \end{aligned}$$

Note that fitting the above models needs no new procedures or software for example, since both cases come under the standard heading. In the first equation all we do is write $\alpha_1 = \beta_1$ and $\alpha_2 = \beta_1 + \beta_2$. In the second we simply redefine the covariates themselves. The equivalence expressed in the above equation is important. It implies two things. Firstly, that this previous question concerning differential treatment effects can be re-expressed in a standard way enabling us to use existing structures, and computer programs. Secondly, since the effects in our models express themselves via products of the form βZ, any recoding of β can be artificially carried out by re-coding Z and vice versa. This turns out to be an important property and anticipates the fact that a non proportional hazards model $\beta(t)Z$ can be re-expressed as a time-dependent proportional hazards model $\beta Z(t)$. Hence the very broad sweep of proportional hazards models.

It is easy to see how the above considerations, applied to a situation in which we have $p > 2$ covariates, become very involved. Suppose we have four ordered levels of some risk factor. We can re-code these levels using three binary covariates as in Table 6.3; For this model we can, again, write the hazard function in terms of these binary coding variables, noting that, as before, there are different ways of expressing this. In standard form we write

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3\}$$

so that the hazard rate for those exposed to the risk factor at level i, $i = 1, \dots, 4$, is given by $\lambda_0(t) \exp(\beta_i)$ where we take $\beta_0 = 0$. Our

Risk factor	Z_1	Z_2	Z_3	Log of risk factor effect
Level 1	0	0	0	0
Level 2	1	0	0	β_1
Level 3	0	1	0	β_2
Level 4	0	0	1	β_3

Table 6.3: Coding for four ordered levels of a risk factor.

interest may be more on the incremental nature of the risk as we increase through the levels of exposure to the risk factor. The above model can be written equivalently as

$$\begin{aligned}\lambda(t|Z) &= \lambda_0(t)\exp\{\beta_1 Z_1 + (\beta_1+\beta_2)Z_2 + (\beta_1+\beta_2+\beta_3)Z_3\}\\ &= \lambda_0(t)\exp\{\beta_1(Z_1+Z_2+Z_3) + \beta_2(Z_2+Z_3) + \beta_3 Z_3\} \quad (6.11)\end{aligned}$$

so that our interpretation of the β_i is in terms of increase in risk. The coefficient β_1 in this formulation corresponds to an overall effect, common to all levels above the lowest. The coefficient β_2 corresponds to the amount by which the log-hazard rate for the second level differs from that at the first. Here then, a value of β_2 equal to zero does not mean that there is no effect at level 2, simply that the effect is no greater than that already quantified at level 1. The same arguments follow for levels 3 and 4.

Writing the model in these different ways is not changing the basic model. It changes the interpretation that we can give to the different coefficients. The equivalent expression shown in Equation 6.11 for example means that we can carefully employ combinations of the covariates in order to use existing software. But we can also consider the original coding of the covariates Z. Suppose that, instead of the coding given in Table (6.3), we use the coding given in Table 6.4. This provides an equivalent description of the four levels. As we move up the levels, changing from level i to level $i+1$, the log hazard is increased by β_i.

Let's imagine a situation, taken from Table 6.4, in which $\beta_1 = \beta_2 = \beta_3$. Real situations may not give rise to strict equalities but may well provide good first approximations. The hazards at each level can now be written very simply as $\lambda_0(t)\exp(j\beta_1)$ for $j = 0, 1, 2, 3$, and this is described in Table (6.5). Taking $\beta_1 = \beta$, we are then able to write a model for this situation as; $\lambda(t|Z) = \lambda_0(t)\exp(\beta Z)$, in which the covariate Z, describing group level, takes the values 0 to 3. This model

Risk factor	Z_1	Z_2	Z_3	Log of risk factor effect
Level 1	0	0	0	0
Level 2	1	0	0	β_1
Level 3	1	1	0	$\beta_1 + \beta_2$
Level 4	1	1	1	$\beta_1 + \beta_2 + \beta_3$

Table 6.4: Coding for four ordered levels of a risk factor.

Risk factor	Z	Log of risk factor effect
Level 1	0	0
Level 2	1	β
Level 3	2	2β
Level 4	3	3β

Table 6.5: Coding for four ordered levels of a risk factor.

has a considerable advantage over the previous one, describing the same situation of four levels, in that only a single coefficient appears in the model as opposed to three. We will use our data to estimate just a single parameter. The gain is clear. The cost, however, is much less so, and is investigated more thoroughly in the chapters on prediction (explained variation, explained randomness) and goodness of fit. If the fit is good, i.e., the assumed linearity is reasonable, then we would certainly prefer the latter model to the former. If we are unsure we may prefer to make less assumptions and use the extra flexibility afforded by a model which includes three binary covariates rather than a single linear covariate. In real data analytic situations we are likely to find ourselves somewhere between the two, using the tools of fit and predictability to guide us.

Returning once more to Table 6.4 we can see that the same idea prevails for the β_i not all assuming the same values. A situation in which four ordered levels is described by three binary covariates could be recoded so that we only have a single covariate Z, together with a single coefficient β. Next, suppose that in the model; $\lambda(t|Z) = \lambda_0(t)\exp(\beta Z)$, Z not only takes the ordered values, 0, 1, 2 and 3 but also all of those in between. In a clinical study this might correspond to some prognostic indicator, such as blood pressure or blood cholesterol, recorded continuously and re-scaled to lie between 0 and 3.

Including the value of Z, as a continuous covariate, in the model amounts to making very strong assumptions. It supposes that the log hazard increases by the same amount for every given increase in Z, so that the relative risk associated with $\Delta = z_2 - z_1$ is the same for all values of z_1 between 0 and $3 - \Delta$. Let's make things a little more involved. Suppose we have the same continuous covariate, this time let's call it Z_1, together with a single binary covariate Z_2 indicating one of two groups. We can write

$$\lambda(t|Z_1, Z_2) = \lambda_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2).$$

Such a model supposes that a given change in exposure Z_1 results in a given change in risk, as just described, but that, furthermore, this resulting change is the same at both levels of the discrete binary covariate Z_2. This may be so but such strong assumptions must be brought under scrutiny. Given the ready availability of software, it is not at all uncommon for data analysts to simply "throw in" all of the variables of interest, both discrete and continuous, without considering potential transformations or recoding, turn the handle, and then try to make sense of the resulting coefficient estimates together with their standard errors. Such an exercise will rarely be fruitful. In this respect it is preferable to write one's own computer programs when possible or to use available software such as the R package, which tends to accompany the user through model development. Packages that present a "complete" one-off black box analysis based on a single model are unlikely to provide much insight into the nature of the mechanisms generating the data at hand.

The user is advised to exercise great care when including continuous covariates in a model. We can view a continuous covariate as equivalent to an infinite dimensional vector of indicator variables so that, in accordance with our informal theorem of complexity, the number of things we need worry about is effectively infinite. Let us not however overstate things, and it is of course useful to model continuous covariates. But be wary. Also consider the model

$$\lambda(t|Z) = \lambda_0(t) \exp(\beta_1 Z + \beta_2 Z^2).$$

If Z is binary then $Z^2 = Z$ and there is no purpose to the second term in the equation. If Z is ordinal or continuous then the effect of Z is quadratic rather than linear. And, adding yet higher-order terms enables us, at least in principle, to model other nonlinear functions. In practice, in order to carry out the analysis, we would use existing tools by simply introducing a second variable Z_2 defined by $Z_2 = Z^2$; an important observation in that the linear representation of the covariate can be relaxed with relatively little effort. For example, suppose that the log-relative risk is expressed via some smooth function $\psi(z)$ of a continuous covariate z. Writing the model

$$\lambda(t|Z) = \lambda_0(t)\exp\{\beta\psi(Z)\}$$

supposes that we know the functional form of the relative risk, at least up to the constant multiple β. Then, a power series approximation to this would allow us to write $\psi(Z) = \sum \beta_j Z^j$ in which any constant term β_0 is absorbed into $\lambda_0(t)$. We then introduce the covariates $Z_j = Z^j$ to bring the model into its standard form.

6.7 Partially proportional hazards models

In the case of a single binary variable, model (6.2) and model (6.4) represent the two extremes of the modeling options open to us. Under model (6.2) there would be no model constraint and any consequent estimation techniques would amount to dealing with each level of the variable independently. Under model (6.4) we make a strong assumption about the nature of the relative hazards, an assumption that allows us to completely share information between the two levels. There exists an important class of models lying between these extremes and, in order to describe this class, let us now imagine a more complex situation; that of three groups, A, B and C, identified by a vector Z of binary covariates; $Z = (Z_2,\ Z_3)$. This is summarized in Table 6.6. We are mainly interested in a treatment indicator Z_1, mindful of the fact that the groups themselves may have very different survival probabilities. Under model (6.4) we have

$$\lambda(t|Z) = \lambda_0(t)\exp\{\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3\}. \tag{6.12}$$

	Z_2	Z_3	Log of group effect
Group A	0	0	0
Group B	1	0	β_2
Group C	1	1	$\beta_2 + \beta_3$

Table 6.6: Coding for three groups.

Our assumptions are becoming stronger in that not only are we modeling the treatment affect via β_1 but also the group effects via β_2 and β_3. Expressing this problem in complete generality, i.e., in terms of model (6.2), we write

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta_1(t)Z_1 + \beta_2(t)Z_2 + \beta_3(t)Z_3\}. \tag{6.13}$$

Unlike the simple case of a single binary variable where our model choices were between the two extremes of model (6.2) and model (6.4), as the situation becomes more complex, we have open to us the possibility of a large number of intermediary models. These are models that make assumptions lying between model (6.2) and model (6.4) and, following O'Quigley and Stare (2002) we call them partially proportional hazards models. A model in between (6.12) and (6.13) is

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta_1 Z_1 + \beta_2(t)Z_2 + \beta_3(t)Z_3\}. \tag{6.14}$$

This model is of quite some interest in that the strongly modeled part of the equation concerns Z_1, possibly the major focus of our study. Figure 6.5 illustrates a simple situation. The only way to leave any state is to die, the probabilities of making this transition varying from state to state and the rates of transition themselves depending on time. Below, under the heading time-dependent covariates, we consider the case where it is possible to move within states. Here it will be possible to move from a low-risk state to a high-risk state, to move from either to the death state, but to also, without having made the transition to the absorbing state, death, to move back from high-risk to low-risk.

Stratified models

Coming under the heading of a partially proportional hazards model is the class of models known as stratified models. In the same way

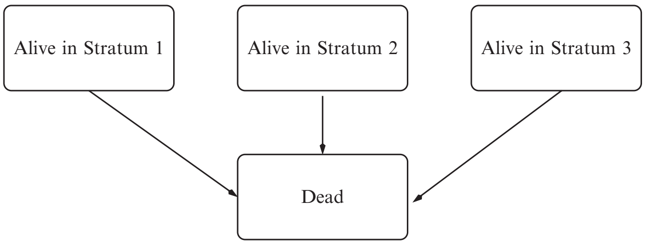

Figure 6.5: A stratified model with transitions only to death state.

these models can be considered as being situated between the two extremes of Equation 6.2 and Equation 6.3 and have been discussed by Kalbfleisch and Prentice (1980) among others. Before outlining why stratified models are simply partially proportional hazards models we recall the usual expression for the stratified model as;

$$\lambda(t|Z(t), w) = \lambda_{0w}(t) \exp\{\beta Z(t)\}, \tag{6.15}$$

where w takes integer values $1, \dots, m$. If the coefficient β were allowed to depend on each stratum, indicated by w, say $\beta(w)$, then this would exactly correspond to a situation in which we consider each stratum completely independent, i.e., we have independent models for each stratum. This would be nothing more than w separate, independent, proportional hazards models. The estimation of $\beta(w)$ for one model has no impact on the estimation of $\beta(w)$ for another. If we take β to be common to the different strata, which is of course the whole purpose of the stratified model, then, using data, whatever we learn about one stratum tell us something about the others. They are no longer independent of one another. Stratified models are necessarily broader than (6.3), lying, in the precise sense described below, between this model and the non proportional hazards model (6.2). To see this, consider a restricted case of model (6.2) in which we have two binary covariates $Z_1(t)$ and $Z_2(t)$. We put the restriction on the coefficient β_2, constrained to be constant in time. The model is then

$$\lambda\{t|Z_1(t), Z_2(t)\} = \lambda_0(t) \exp\{\beta_1(t) Z_1(t) + \beta_2 Z_2(t)\}, \tag{6.16}$$

a model clearly lying, in a well-defined way, between models (6.3) and (6.2). It follows that

$$\lambda\{t|Z_1(t) = 0, Z_2(t)\} = \lambda_0(t) \exp\{\beta_2 Z_2(t)\}$$

and

$$\lambda\{t|Z_1(t) = 1, Z_2(t)\} = \lambda_0^*(t) \exp\{\beta_2 Z_2(t)\},$$

where $\lambda_0^*(t) = \lambda_0(t) e^{\beta_1(t)}$. Recoding the binary $Z_1(t)$ to take the values 1 and 2, and rewriting $\lambda_0^*(t) = \lambda_{02}(t)$, $\lambda_0(t) = \lambda_{01}(t)$ we recover the stratified PH model (6.15) for $Z_2(t)$. The argument is easily seen to be reversible and readily extended to higher dimensions so we can conclude an equivalence between the stratified model and the partially proportional hazards model in which some of the $\beta(t)$ are constrained to be constant. We can exploit this idea in the goodness of fit or the model construction context. If a PH model holds as a good approximation, then the main effect of Z_2 say, quantified by β_2, would be similar over different stratifications of Z_1 and remain so when these stratifications are re-expressed as a PH component to a two covariate model. Otherwise the indication is that $\beta_1(t)$ should be allowed to depend on t. The predictability of any model is studied later under the headings of explained variation and explained randomness and it is of interest to compare the predictability of a stratified model and an un-stratified one. For instance, we might ask ourselves just how strong is the predictive strength of Z_2 after having accounted for Z_1. Since we can account for the effects of Z_1 either by stratification or by its inclusion in a single PH model we may obtain different results. Possible discrepancies tell us something about our model choice.

The relation between the hazard function and the survival function follows as a straightforward extension of (6.5). Specifically, we have

$$S(t|Z) = \sum_w \phi(w)\{S_{0w}(t)\}^{\exp(\beta Z)}, \tag{6.17}$$

where $S_{0w}(t)$ is the corresponding baseline survival function in stratum w and $\phi(w)$ is the probability of coming from that particular stratum. This is then slightly more involved than the nonstratified case in which, for two groups the model expressed the survival function of one group as a power transformation of the other. Nonetheless the connection to the class of Lehmann alternatives is still there although somewhat weaker. For the stratified model, once again the quantity $\lambda_{0w}(t)$ does not appear in the expression for the partial likelihood given now by

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta Z_i)}{\sum_{j=1}^{n} Y_j\{w_i(X_i), X_i\} \exp(\beta Z_j)} \right\}^{\delta_i} \tag{6.18}$$

and, in consequence, once again, $\lambda_{0w}(t)$ can remain arbitrary. Note also that each term in the product is the conditional probability that at time X_i of an observed failure, it is precisely individual i who is selected to fail, given all the individuals at risk from stratum w and that one failure from this stratum occurs.

The notation $w_i(t)$ indicates the stratum in which the subject i is found at time t. Although we mostly consider $w_i(t)$ which do not depend on time, i.e., the stratum is fixed at the outset and thereafter remains the same, it is almost immediate to generalize this idea to time dependency and we can anticipate the later section on time-dependent covariates where the risk indicator $Y_j\{w_i(t),t\}$ is not just a function taking the value one until it drops at some point to zero, but can change between zero and one with time, as the subject moves from one stratum to another. For now the function $Y_j\{w_i(t),t\}$ will be zero unless the subject is at risk of failure from stratum w_i, i.e., the same stratum in which the subject i is to be found. Taking the logarithm in (6.18) and derivative with respect to β, we obtain the score function

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left\{ Z_i - \frac{\sum_{j=1}^{n} Y_j\{w_i(X_i), X_i\} Z_j \exp(\beta Z_j)}{\sum_{j=1}^{n} Y_j\{w_i(X_i), X_i\} \exp(\beta Z_j)} \right\}, \quad (6.19)$$

which, upon setting equal to zero, can generally be solved without difficulty using standard numerical routines, to obtain the maximum partial likelihood estimate $\hat{\beta}$ of β. The parameter β then is assumed to be common across the different strata.

Inferences about β are made by treating $\hat{\beta}$ as asymptotically normally distributed with mean β and variance $I(\hat{\beta})^{-1}$, where, now, $I(\beta)$ is given by $I(\beta) = \sum_{i=1}^{n} \delta_i I_i(\beta)$. In this case each I_i is, as before, obtained as the derivative of each component to the score statistic $U(\beta)$. For the stratified score this is

$$\begin{aligned} I_i \;=\; & \frac{\sum_{j=1}^{n} Y_j\{w_i(X_i), X_i\} Z_j^2 \exp(\beta Z_j)}{\sum_{j=1}^{n} Y_j\{w_i(X_i), X_i\} \exp(\beta Z_j)} \\ & - \left\{ \frac{\sum_{j=1}^{n} Y_j\{w_i(X_i), X_i\} Z_j \exp(\beta Z_j)}{\sum_{j=1}^{n} Y_j\{w_i(X_i), X_i\} \exp(\beta Z_j)} \right\}^2 . \end{aligned}$$

The central notion of the risk set is once more clear from the above expressions and we most usefully view the score function as contrasting the observed covariates at each distinct failure time with the means of those at risk from the same stratum. A further way of looking at the score function is to see it as having put the individual contributions on a linear scale. We simply add them up within a stratum and then, across the strata, it only remains to add up the different sums. Once again, inferences can also be based on likelihood ratio methods or on the score $U(\beta)$, which in large samples can be considered to be normally distributed with mean zero and variance $I(\beta)$. Multivariate extensions follow as before. For the stratified model the only important distinction impacting the calculation of $U(\beta)$ and $I_i(\beta)$ is that the sums are carried out over each stratum separately and then combined at the end. The indicator $Y_j\{w_i(X_i)\}$ enables this to be carried out in a simpler way as indicated by the equation.

Random effects and frailty models

Also coming under the heading of partially proportional hazards model are the classes of models, which include random effects. When the effects concern a single individual such models have been given the heading frailty models (Vaupel 1979) since, for an individual identified by w, we can write $\lambda_{0w}(t) = \alpha_w \lambda_0(t)$ implying a common underlying hazard $\lambda_0(t)$ adjusted to each individual by a factor, the individual's *frailty*, unrelated to the effects of any other covariates that are quantified by the regression coefficients. The individual effects are then quantified by the α_w.

Although of some conceptual interest, such models are indistinguishable from models with time-dependent regression effects and therefore, unless there is some compelling reason to believe (in the absence of frailties) that a proportional hazards model would hold, it seems more useful to consider departures from proportional hazards in terms of model (6.2). On the other hand, random effects models, as commonly described by Equation (6.20) in which the α_w identify a potentially large number of different groups, are interesting and potentially of use. We express these as

$$\lambda(t|Z(t), w) = \alpha_w \lambda_0(t) \exp\{\beta Z(t)\}. \tag{6.20}$$

These models are also partially parametric in that some effects are allowed not to follow a proportional hazards constraint. However, unlike the stratified models described above, restrictions are imposed. The most useful view of a random effects model is to see it as a stratified model with some structure imposed upon the strata. A random effects model is usually written

$$\lambda(t|Z(t), w) = \lambda_0(t) \exp\{\beta Z(t) + w\}, \tag{6.21}$$

in which we take w as having been sampled from some distribution $G(w; \theta)$. Practically there will only be a finite number of distinct values of w, however large. For any value w we can rewrite $\lambda_0(t)e^w = \lambda_{0w}(t)$ and recover model (6.15). For the right hand side of this equation, and as we might understand from (6.15), we suppose w to take the values 1,2, ... The values on the left-hand side, being generated from

$G(\cdot)$ would generally not be integers but this is an insignificant notational issue and not one involving concepts. Consider the equation to hold. It implies that the random effects model is a stratified model in which added structure is placed on the strata. In view of Equation 6.16 and the arguments following this equation we can view a random effects model equivalently as in Equation 6.2 where, not only are PH restrictions imposed on some of the components of $\beta(t)$, but the time dependency of the other components is subject to constraints. These latter contraints, although weaker than imposing constancy of effect, are all the stronger as the distribution of $G(w;\theta)$ is concentrated.

Structure of random effects models

Consider firstly the model of Equation (6.3). Suppose we have one main variable, possibly a treatment variable of interest, coded by $Z_1 = 0$ for group A and $Z_1 = 1$ for group B. The second variable, say a center variable, which may or may not have prognostic importance and for which we may wish to control for possible imbalance is denoted Z_2. A strong modeling approach would include both binary terms in the model so that the relationship between the hazard functions is as described in Figure 6.6. If our main focus is on the effect of treatment, believed to be comparable from one center to another, even though the effects of the centers themselves are not absent, it makes sense to stratify. This means that we do not attempt to model the effects of the centers but, instead, remove any such potential effects from our

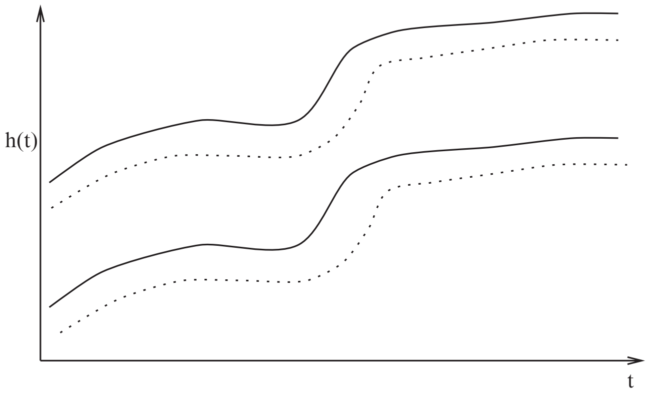

Figure 6.6: PH model with binary covariates denoting center and treatment groups.

analysis. This is nice in that it allows for rather greater generality than that illustrated in Figure 6.6. We maintain an assumption of constant treatment effect but the center effects can be arbitrary. This is illustrated in Figure 6.7. The illustration makes it clear that, under the assumption, a weaker one than that implied by Equation 6.3, we can estimate the treatment effect whilst ignoring center effects. A study of these figures is important to understanding what takes place when we impose a random effects model as in Equation (6.21). For many centers, Figure 6.8, rather than having two curves per center, parallel but otherwise arbitrary, we have a family of parallel curves. We no

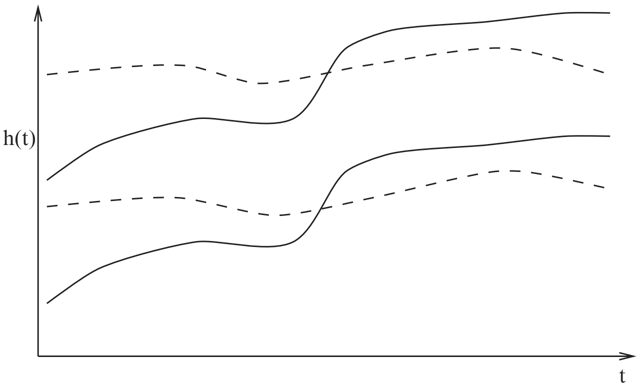

Figure 6.7: An outline sketch of a stratified PH model. Main variable in two strata: stratum 1, i.e., center 1 given by dotted line; stratum 2 by continuous line.

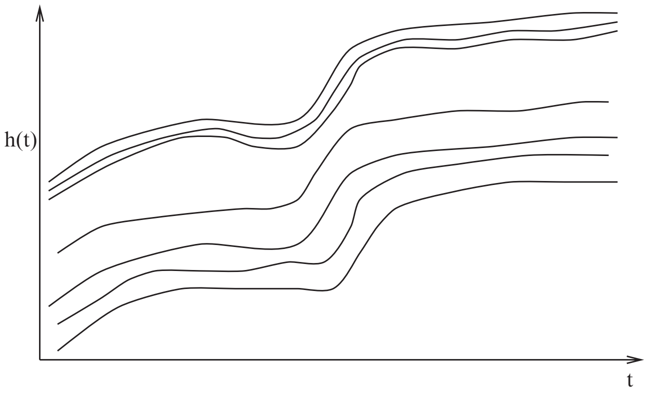

Figure 6.8: An outline sketch of a PH model with centers as random effects.

longer are able to say anything about the distance between any given centers, as we could for the model of Equation 6.3, a so-called fixed effects model, but the distribution of the distances between centers is something we aim to quantify. This is summarized by the distribution $G(w;\theta)$ and our inferences are then partly directed at θ.

Random effects models versus stratified models

The stratified model is making weaker assumptions than the random effects model. This follows since the random effects model is just a special case of a stratified model in which some structure is imposed upon the differences between strata. The stratified model not only leaves any distribution of differences between strata unspecified, but it also makes no assumption about the form of any given stratum. Whenever the stratified model is valid, then so also is the random effects model, the converse not being the case.

It may then be argued that we are making quite a strong assumption when we impose this added structure upon the stratified model. In exchange we would hope to make non-negligible inferential gains, i.e., greater precision of our estimates of errors for the parameters of main interest, the treatment parameters. In practice gains tend to be small for most situations and give relatively little reward for the extra effort made. Since any such gains are only obtainable under the assumption that the chosen random effects model actually generates the data, actual gains in practice are likely to be yet smaller and, of course, possibly negative when our additional model assumptions are incorrect. A situation where gains for the random effects model may be of importance is one where a non-negligeable subset of the data include strata containing only a single subject. In such a case simple stratification would lose information on those subjects. A random effects model, assuming the approximation to be sufficiently accurate, enables us to recover such information.

Efficiency of random effects models

Most of our discussion here focuses on different possible representations of the infinitely complex reality we are hoping to model. Our purpose in modeling is, ultimately, to draw simple, at least clear-cut, inferences. The question of inference no longer concerns the general but rather the specific data set we have at hand. If our main concern is on estimating risk functions then the question becomes, to what extent do we gain by including in our inferential setup the presence of random effect terms. Since our main objective is estimation and quantification of regression parameters enabling us to say something about the risk factors under study, the idea behind the inclusion of additional random effect terms is to make more precise this estimation and quantification.

As already argued above the inclusion of individual random effects (frailties) is of no practical interest and simply amounts to expressing the idea, albeit in an indirect way, of model inadequacy (O'Quigley and Stare 2002). We therefore assume that we are dealing with groups, some of which, but not all, may only include a isolated individual. We know that a partial likelihood analysis, stratified by group, is estimating the same regression parameter. Inference is based on the stratified score statistic. We contrast the observed covariate value with its estimated expectation under the model. Different model assumptions will impact this estimated expectation and it is here that any efficiency gains can be made. For a stratified model, these estimated expectations may be with respect to relatively small risk sets. A random effects model on the other hand, via the inclusion of a different w per group, will estimate the relevant expectations over the whole risk set and not just that relative to the group defined by the covariate value.

Comparisons for the stratified model are made with respect to the relatively few subjects of the group risk sets. This may lead us to believe that much information could be recovered were we able to make the comparison, as does the alternative random effects analysis, with respect to the whole risk set. Unfortunately this is not quite so because each contribution to the score statistic involves a difference between an observation on a covariate and its expectation under the model and the "noise" in the expectation estimate is of lower order that the covariate observations themselves. There is not all that much to be gained by improving the precision of the expectation estimate.

In other words, using the whole of the risk set or just a small sample from it will provide similar results. This idea of risk set sampling has been studied in epidemiology and it can be readily seen that the efficiency of estimates based on risk set samples of size k, rather than the whole risk set, is of the order

$$\frac{k}{k+1}\left\{1+\sum_{j=1}^{n}\frac{1}{n(n-j+1)}\right\}. \tag{6.22}$$

This function increases very slowly to one but, with as few as four subjects on average in each risk set comparison, we have already achieved 80% efficiency. With nine subjects this figure is close to 90%. Real efficiency will be higher for two reasons: (1) the above assumes that the estimate based on the full risk set is without error, (2) in our context we are assuming that each random effect w is observed precisely.

Added to this is the fact that, since the stronger assumptions of the random effects model must necessarily depart to some degree from the truth, it is by no means clear that there is much room to make any kind of significant gains. As an aside, it is of interest to note that, since we do not gain much by considering the whole of the risk set as opposed to a small sample from it, the converse must also hold, i.e., we do not lose very much by working with small samples rather than the whole of the risk set. In certain studies, there may be great economical savings made by only using covariate information, in particular when time dependent, from a subset of the full risk set.

Table 6.7 was taken from O'Quigley and Stare (2002). The table was constructed from simulated failure times where the random effects model was taken to be exactly correct. Data were generated from this model in which the gamma frailty had a mean and variance equal to one. The regression coefficient of interest was exactly equal to 1.0. Three situations were considered; 100 strata each of size 5, 250 strata each of size 2 and 25 strata each of size 20. The take-home message from the table is that, in these cases for random effects models, not much is to be gained in terms of efficiency. Any biases appear negligible and the mean of the point estimates for both random effects and stratified models, while differing notably from a crude model ignoring model inadequacy, are effectively indistinguishable. As we would expect there is a gain for the variance of estimates based on the random effects model but, even for highly stratified data (100 × 5), any gain is very small. Indeed for the extreme case of 250 strata, each of size 2, surely the worst situation for the stratified model, it is difficult to become enthusiastic over the comparative performance of the random effects model.

	100×5	250×2	25×20
Ignoring effect	0.52 (0.16)	0.51 (0.16)	0.54 (0.16)
Random effect model	1.03 (0.19)	0.99 (0.22)	1.01 (0.17)
Stratified model	1.03 (0.22)	1.02 (0.33)	1.01 (0.18)

Table 6.7: Simulations for three models under different groupings.

We might conclude that we only require around 80% of the comparative sample size needed for estimating relative risk based on the stratified model. But, such a conclusion, leaning entirely on the assumption that we know not only the class of distributions from which the random effects come but also the exact value of the population parameters, suggests, in practice, that the hoped for gain, in this most hopeful of cases, is more likely to be greater than the 80% indicated by our calculations. The only real situation that can be clearly disadvantageous to the stratified model is one where a non-negligible subset of the strata are seen to only contain a single observation. For such cases, and assuming a random effects model to provide an adequate fit, information from states with a single observation (which would be lost by a stratified analysis) can be recovered by a random effects analysis.

6.8 Non proportional hazards model with intercept

Recalling the general model, i.e., the non proportional hazards model for which there is no restriction on $\beta(t)$, note that we can re-express this so that the function $\beta(t)$ is written as a constant term, the intercept, plus some function of time multiplied by a constant coefficient. Writing this as

$$\lambda(t|Z) = \lambda_0(t) \exp\{[\beta_0 + \theta Q(t)]Z\}, \tag{6.23}$$

we can describe the term β_0 as the intercept and $Q(t)$ as reflecting the nature of the time dependency. The coefficient θ will simply scale this dependency and we may often be interested in testing the particular value, $\theta = 0$, since this value corresponds to a hypothesis of proportional hazards. Fixing the function $Q(t)$ to be of some special functional form allows us to obtain tests of proportionality against alternatives of a particular nature. Linear or quadratic decline in the log-relative risk, change-point, and crossing hazard situations are all then easily accommodated by this simple formulation. Tests of goodness of fit of the proportional hazards assumption can be then be constructed which may be optimal for certain kinds of departures.

Although not always needed it can sometimes be helpful to divide the time axis into r nonoverlapping intervals, $B_1, \ldots, B_r$ in an ordered sequence beginning at the origin. In a data-driven situation these intervals may be chosen so as to have a comparable number of events in each interval or so as not to have too few events in any given interval. Defined on these intervals is a vector, also of dimension r, of some known or estimable functions of time, not involving the parameters of interest, β. This is denoted $Q(t) = \{Q_1(t), \ldots, Q_r(t)\}$ This model is then written in the form,

$$\lambda(t|Z) = \lambda_0(t) \exp\{[\beta + \theta Q(t)]Z\}, \tag{6.24}$$

where θ is a vector of dimension r. Thus, $\theta Q(t)$ (here the usual inner product) has the same dimension as β, i.e., one. In order to investigate the time dependency of particular covariates in the case of multivariate Z we would have β of dimension greater than one, in which case $Q(t)$ and θ are best expressed in matrix notation (O'Quigley and Pessione 1989).

Here, as through most of this text, we concentrate on the univariate case since the added complexity of the multivariate notation does not bring any added light to the concepts being discussed. Also, for the majority of the cases of interest, $r = 1$ and θ becomes a simple scalar. We will often have in mind some particular form for the time-dependent regression coefficient $Q(t)$, common examples being a linear slope (Cox 1972), an exponential slope corresponding to rapidly declining effects (Gore et al. 1984) or some function related to the marginal distribution, $F(t)$ (Breslow, Edler and Berger 1984). In practice we may be able to estimate this function of $F(t)$ with the help of consistent estimates of $F(t)$ itself, in particular the Kaplan-Meier estimate. The non proportional hazards model with intercept is of particular use in questions of goodness of fit of the proportional hazards model pitted against specific alternatives. These specific alternatives can be quantified by appropriate forms of the function $Q(t)$. We could also test a joint null hypothesis $H_0 : \beta = \theta = 0$ corresponding to no effect, against an alternative H_1, either θ or β nonzero. This leads to a test with the ability to detect non proportional hazards, as well as proportional hazards departures to the null hypothesis of no effect. We could also test a null hypothesis $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$, leaving β itself unspecified. This would then provide a goodness-of-fit test of the proportional hazards assumption. We return to these issues later on when we investigate in greater detail how these models give rise to simple goodness of fit tests.

Changepoint models

A simple special case of a non proportional hazards model with an intercept is that of a changepoint model. O'Quigley and Pessione (1991), O'Quigley (1994), and O'Quigley and Natarajan (2004) develop such models whereby we take the function $Q(t)$ to be defined by, $Q(t) = I(t \leq \gamma) - I(t > \gamma)$ with γ an unknown changepoint. This function $Q(t)$ depends upon γ but otherwise does not depend upon the unknown regression coefficients and comes under the above heading of a non proportional hazards model with an intercept. For the purposes of a particular structure for a goodness of fit test we can choose the intercept to be equal to some fixed value, often zero (O'Quigley and Pessione 1991). The model is then

$$\lambda(t|Z) = \lambda_0(t) \exp\{[\beta + \alpha Q(t)]Z(t)\}. \tag{6.25}$$

The parameter α is simply providing a scaling (possibly of value zero) to the time dependency as quantified by the function $Q(t)$. The chosen form of $Q(t)$, itself fixed and not a parameter, determines the way in which effects change through time; for instance whether they decline exponentially to zero, whether they decline less rapidly or any other way in which effects might potentially change through time.

Inference for the changepoint model is not straightforward and in the series of chapters dealing with approaches to inference one chapter is devoted specifically to changepoint models. Note that were γ to be known, then inference would come under the usual headings with no additional difficulty. The changepoint model expressed by Equation (6.25) deals with the regression effect changing through time and putting the model under the heading of a non proportional hazards model. A related, although entirely different model, is one which arises as a simplification of a proportional model with a continuous covariate and the idea is to replace the continuous covariate by a discrete classification.

The classification problem itself fall into two categories. If we are convinced of the presence of effects and simply wish to derive the most predictive classification into, say, two groups, then the methods using explained randomness or explained variation will achieve this goal. If, on the other hand, we wish to test a null hypothesis of absence of effect, and, in so doing, wish to consider all possible classifications based on a family of potential cutpoints of the continuous covariate, then special techniques of inference are required. We return to this in Chapter 12.

6.9 Time-dependent covariates

In all of the above models we can make a simple change by writing the covariate Z as $Z(t)$, allowing the covariate to assume different values at different time points. Our model then becomes

$$\lambda(t|Z(t)) = \lambda_0(t)\exp\{\beta(t)Z(t)\} \tag{6.26}$$

and allows situations such as those described in Figure 6.9 to be addressed. As we change states the intensity function changes. This enables us to immediately introduce further refinement into a simple alive/dead model whereby we can suppose one or more intermediary states. A subject can move across states thereby allowing prognosis to improve or to worsen, the rates of these changes themselves depending upon other factors. The state death is described as an absorbing state and so we can move into this state but, once there, we cannot move out of it again.

Mostly we will work with the proportional hazard restriction on the above model so that

$$\lambda(t|Z(t)) = \lambda_0(t)\exp\{\beta Z(t)\}, \tag{6.27}$$

Such a simple, albeit very much more sophisticated, model than our earlier one describes a broad range of realistic situations. We will see that models with time-dependent covariates do not raise particular difficulties, either computationally or from the viewpoint of interpretation, when we deal with inference. This will be clear from the main

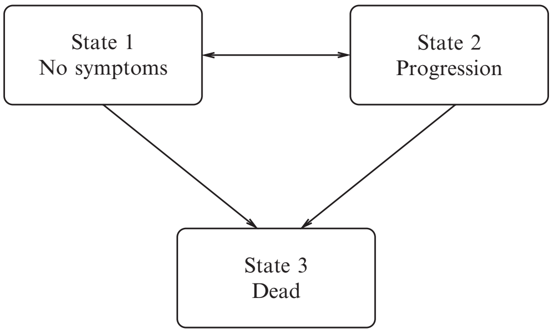

Figure 6.9: Compartment model where ability to move between states other than death state can be characterized by time dependent indicator covariates $Z(t)$.

theorem of proportional hazards regression (Section 7.4). The model simply says that the effect of the covariate remains constant, i.e., the regression coefficient remains constant, but that the covariate, or state, can itself change with time. Models with time-dependent covariates can also be used as a purely artificial construction in order to be able to express non proportional hazards models in a proportional hazards form. That is not our main purpose here, however, and we are assuming that $Z(t)$ does correspond to some real physical measurement which can be obtained through time.

We can also imagine a slightly more involved situation than the above. Suppose that the covariate Z remains fixed, but that a second covariate, known to influence survival, also needs to be accounted for. Furthermore this second covariate is time dependent. We could, of course, simply use the above model extended to the case of two covariates. This is straightforward, apart from the fact that, as previously underlined by the complexity theorem, care is needed. If, however, we do not wish to model the effects of this second covariate, either because it is only of indirect concern or because its effects might be hard to model, then we could appeal to a stratified model. We write;

$$\lambda(t|Z(t), w(t)) = \lambda_{0w(t)}(t) \exp\{\beta Z(t)\}, \tag{6.28}$$

where, as for the non time-dependent case, $w(t)$ takes integer values 1,..., m indicating status. The subject can move in and out of the m strata as time proceeds. Two examples illustrate this. Consider a new treatment to reduce the incidence of breast cancer. An important time-dependent covariate would be the number of previous incidents of benign disease. In the context of inference, the above model simply means that, as far as treatment is concerned, the new treatment and the standard are only ever contrasted within patients having the same previous history. These contrasts are then summarized in final estimates and possibly tests. Any patient works her way through the various states, being unable to return to a previous state. The states themselves are not modeled. A second example might be a sociological study on the incidence of job loss and how it relates to covariates of main interest such as training, computer skills etc. Here, a stratification variable would be the type of work or industry in which the individual finds him or herself. Unlike the previous example a subject can move between states and return to previously occupied states.

Time-dependent covariates describing states can be used in the same way for transition models in which there is more than one

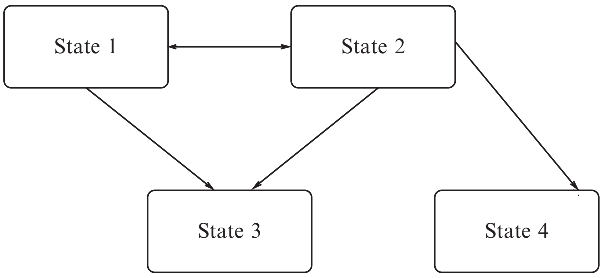

Figure 6.10: Compartment model with 2 absorbing "death" states.

absorbing "death" state. Many different kinds of situations can be constructed, these situations being well described by compartment models with arrows indicating the nature of the transitions that are possible (Figure 6.10). For compartment models with time-dependent covariates there is a need for some thought when our interest focuses on the survival function. The term external covariate is used to describe any covariate $Z(t)$ such that, at $t = 0$, for all other $t > 0$, we know the value of $Z(t)$. The paths can be described as deterministic. In the great majority of the problems that we face this is not the case and a more realistic way of describing the situation is to consider the covariate path $Z(t)$ to be random. Also open to us as a modeling possibility, when some covariate $Z_1(t)$ is of secondary interest assuming a finite number of possible states, is to use the at risk function $Y(s,t)$. This restricts our summations to those subjects in state s as described above for stratified models.

6.10 Time-dependent covariates and non proportional hazards models

A non proportional hazard model with a single constant covariate Z is written

$$\lambda(t|Z) = \lambda_0(t)\exp\{\beta(t)Z\}. \qquad (6.29)$$

The multivariate extension is immediate and, in keeping with our convention of only dealing with the univariate problem whenever possible, we focus our attention on the simple product $\beta(t)Z$ at some given point in time t. If we define $\beta_0 = \beta(0)$ we can rewrite this product as $\beta_0 Z(t)$ where $Z(t) = Z\beta(t)/\beta_0$. We could take any other time point t'

and, once again, we observe that we can re-write the product $\beta(t')Z$ as $\beta_0 Z(t')$ where $Z(t') = Z\beta(t')/\beta_0$. This equivalence is then true for any, and all, values of t. Thus, a non proportional hazards model with a constant covariate can be re-expressed, equivalently, as a simple proportional hazards model with a time-dependent covariate.

It is almost immediate, and perhaps worth carrying out as an exercise, to show that we can reverse these steps to conclude also that any model with time dependent covariates can be expressed in an equivalent form as a non proportional hazards model. In conclusion, for every non proportional hazards model there exists an equivalent proportional hazards model with time-dependent covariates. Indeed, it also clear that this argument can be extended. For, if we have the model; $\lambda(t|Z) = \lambda_0(t)\exp\{\beta(t)Z(t)\}$, then, via a re-expression of the model using $\beta_0 Z^*(t)$ where $Z^*(t) = Z(t)\beta(t)/\beta_0$, we can construct a proportional hazards model with a time-dependent regression effect from a model which began with both time-dependent regression effects as well as time changing regression coefficient.

This equivalence is a formal one and does not of itself provide any new angle on model development. It may be exploited nonetheless in theoretical investigation or used as a means to enable the structuring of a particular problem. For example, many available softwares, as well as user written code, will cater for time-dependent covariables. This facility can then be made use of should we wish to study particular types of non proportional hazards models.

6.11 Proportional hazards models in epidemiology

For arbitrary random variables X and Y with joint density $f(x, y)$, conditional densities $g(x|y)$ and $h(y|x)$, marginal densities $v(x)$ and $w(y)$, we know that

$$f(x, y) = g(x|y)w(y) = h(y|x)v(x),$$

so that, in the context of postulating a model for the pair (X, Y), we see that there are two natural potential characterizations. Recalling the discussion from Section 4.3 note that, for survival studies, our interest in the binary pair (T, Z), time and covariate, can be seen equivalently from the viewpoint of the conditional distribution of time given the covariate, along with the marginal distribution of the covariate, or from the viewpoint of the conditional distribution of the covariate given time, along with the marginal distribution of time. This equivalence we exploit in setting up inference where, even though the physical problem concerns time given the covariate, our analysis describes the distribution of the covariate given time.

In epidemiological studies the variable time T is typically taken to be age. Calendar time and time elapsed from some origin may also be used but, mostly, the purpose is to control for age in any comparisons we wish to make. Usually we will consider rates of incidence of some disease within small age groups or possibly, via the use of models, for a large range of values of age. Unlike the relatively artificial construction of survival analysis which exploits the equivalent ways of expressing joint distributions, in epidemiological studies our interest naturally falls on the rates of incidence for different values of Z given fixed values of age T. It is not then surprising that the estimating equations we work with turn out to be essentially the same for the two situations.

The main theorem of proportional hazards regression (Section 7.4) applies more immediately in epidemiology than in survival type studies. We return to this in the chapter on inference. One important distinction, although already well catered for by use of our "at risk" indicator variables, is that for epidemiological studies the subjects in different risk sets are often distinct subjects. Even so, as we will see, the form of the equations is the same, and software which allows an analysis of survival data will also allow an analysis of certain problems in epidemiology.

For a binary outcome, indicated by $Y = 1$ or $Y = 0$, and a binary risk or exposure factor, $Z = 1$ or $Z = 0$, the relative risk is defined as the ratio of the probabilities $P(Y = 1|Z = 1)/P(Y = 1|Z = 0)$ and the, related, odds ratio ψ as

$$\psi = \frac{P(Y = 1|Z = 1)P(Y = 0|Z = 0)}{P(Y = 1|Z = 0)P(Y = 0|Z = 1).}$$

In the above and in what follows, in order for the notation not to become too cluttered, we write $\Pr(A) = P(A)$. Under a "rare disease assumption," i.e., when $P(Y = 0|Z = 0)$ and $P(Y = 0|Z = 1)$ are close to 1, then the odds ratio and relative risk approximate one another. One reason for being interested in the odds ratio, as a measure of the impact of different levels of the covariate (risk factor) Z follows from the easily obtained identity

$$\frac{P(Y=1|Z=1)P(Y=0|Z=0)}{P(Y=1|Z=0)P(Y=0|Z=1)} = \frac{P(Z=1|Y=1)P(Z=0|Y=0)}{P(Z=1|Y=0)P(Z=0|Y=1)}. \tag{6.30}$$

Thus, the impact of different levels of the risk factor Z can equally well be estimated by studying groups defined on the basis of this same risk factor and their corresponding incidence rates of $Y = 1$. This provides the rationale for the case-control study in which, in order to estimate ψ, we make our observations on Z over fixed groups of cases and controls (distribution of Y fixed), rather than the more natural, but practically difficult if not impossible, approach of making our observations on Y for a fixed distribution of Z. Assumptions and various subtleties are involved. The subject is vast and we will not dig too deeply into this. The points we wish to underline in this section are those that establish the link between epidemiological modeling and proportional hazards regression.

Series of 2 × 2 *tables*

The most elementary presentation of data arising from either a prospective study (distribution of Z fixed) or a case-control study (distribution of Y fixed) is in the form of a 2×2 contingency table in which the counts of the number of observations are expressed. Estimated probabilities, or proportions of interest are readily calculated. In Table 6.8, $a_{1*} = a_{11} + a_{12}$, $a_{2*} = a_{21} + a_{22}$, $a_{*1} = a_{11} + a_{21}$, $a_{*2} = a_{12} + a_{22}$ and $a_{**} = a_{1*} + a_{2*} = a_{*1} + a_{*2}$. For prospective studies the proportions a_{11}/a_{*1} and a_{12}/a_{*2} estimate the probabilities of being a case ($Y = 1$) for both exposure groups while, for case-control studies, the proportions a_{11}/a_{1*} and a_{21}/a_{2*} estimate the

	$Z = 1$	$Z = 0$	totals
$Y = 1$	a_{11}	a_{12}	a_{1*}
$Y = 0$	a_{21}	a_{22}	a_{2*}
Totals	a_{*1}	a_{*2}	a_{**}

Table 6.8: Basic 2×2 table for cases ($Y = 1$) and controls ($Y = 0$).

probabilities of exhibiting the risk or exposure factor ($Z = 1$) for both cases and controls. For both types of studies we can estimate ψ by the ratio $(a_{11}a_{22})/(a_{21}a_{12})$, which is also the numerator of the usual chi-squared test for equality of the two probabilities. If we reject the null hypothesis of the equality of the two probabilities we may wish to say something about how different they are based on the data from the table. As explained below, under the heading "Logistic regression," quantifying the difference between two proportions is not best done via the most obvious, and simple, arithmetic difference. There is room for more than one approach, the simple arithmetic difference being perfectly acceptable when sample sizes are large enough to be able to use the De Moivre-Laplace approximation (Section 3.3) but, more generally, the most logical in our context is to express everything in terms of the odds ratio. We can then exploit the following theorem;

Theorem 6.2 *Taking all the marginal totals as fixed, the conditional distribution of* a_{11} *is written*

$$P(a|a_{1*}, a_{2*}, a_{*1}, a_{*2}) = \binom{a_{1*}}{a}\binom{a_{2*}}{a_{*1} - a}\psi^a \Big/ \sum_u \binom{a_{1*}}{u}\binom{a_{2*}}{a_{*1} - u}\psi^u,$$

the sum over u being over all integers compatible with the marginal totals. The conditionality principle appears once more, in this instance in the form of fixed margins. The appropriateness of such conditioning, as in other cases, can be open to discussion. But again, insightful conditioning has greatly simplified the inferential structure. Following conditioning of the margins, it is only necessary to study the distribution of any one entry in the 2×2 table, the other entries being then determined. It is usual to study the distribution of a_{11}. A nonlinear estimating equation can be based on $a_{11} - E(a_{11})$, expectation obtained from Theorem 6.2, and from which we can estimate ψ and associate a variance term with the estimator. The nonlinearity of the estimating equation, the only approximate normality of the estimator, and the involved form of variance expressions has led to much work in the methodological epidemiology literature; improving the approximations, obtaining greater robustness and so on. However, all of this can be dealt with in the context of a proportional hazards (conditional logistic) regression model. Since it would seem more satisfactory to work with a single structure rather than deal with problems on a case-by-case basis the recommendation is to work with proportional and non proportional hazards models. Not only does a model enable us

Table i	$Z = 1$	$Z = 0$	Totals
$Y = 1$	$a_{11}(i)$	$a_{12}(i)$	$a_{1*}(i)$
$Y = 0$	$a_{21}(i)$	$a_{22}(i)$	$a_{2*}(i)$
Totals	$a_{*1}(i)$	$a_{*2}(i)$	$a_{**}(i)$

Table 6.9: 2×2 table for ith age group of cases and controls.

to more succinctly express the several assumptions which we may be making it offers, more readily, well established ways of investigating the validity of any such assumptions. In addition the framework for studying questions such as explained variation, explained randomness and partial measures of these is clear and requires no new work.

The "rare disease" assumption, allowing the odds ratio and relative risk to approximate one another, is not necessary in general. However, the assumption can be made to hold quite easily and is therefore not restrictive. To do this we construct fine strata, within which the probabilities $P(Y = 0|Z = 0)$ and $P(Y = 0|Z = 1)$ can be taken to be close to 1. For each stratum, or table, we have a 2×2 table as in Table 6.9, indexed by i. Each table provides an estimate of relative risk at that stratum level and, assuming that the relative risk itself does not depend upon this stratum, although the actual probabilities themselves composing the relative risk definition may themselves depend upon strata, then the problem is putting all these estimates of the same thing into a single expression. The most common such expression for this purpose is the Mantel-Haenszel estimate of relative risk.

Mantel-Haenszel estimate of relative risk

The, now famous, Mantel-Haenszel estimate of relative risk was described by Mantel and Haenszel (1959) and is particularly simple to calculate. Referring to the entries of observed counts in Table 6.10, if we first define for the i th subtable $R_i = a_{11}(i)a_{22}(i)/a_{**}(i)$ and $S_i = a_{12}(i)a_{21}(i)/a_{**}(i)$, then the Mantel-Haenszel summary relative risk estimate across the tables is given by $\hat{\psi}_{MH} = \sum_i R_i / \sum_i S_i$. Breslow (1996) makes the following useful observations concerning $\hat{\psi}_{MH}$ and $\hat{\beta}_{MH} = \hat{\psi}_{MH}$. First, $E(R_i) = \psi_i E(S_i)$ where the true odds ratio in the ith table is given by ψ_i. When all of these odds ratios coincide then $\hat{\psi}_{MH}$ is the solution to the unbiased estimating equation; $R - \psi S = 0$, where $R = \sum_i R_i$ and $S = \sum_i S_i$. Under an assumption of

Table i	$Z=1$	$Z=0$	Totals
$Y=1$	$a_{11}(i)$	$a_{12}(i)$	$a_{1*}(i)$
$Y=0$	$a_{21}(i)$	$a_{22}(i)$	$a_{2*}(i)$
Totals	$a_{*1}(i)$	$a_{*2}(i)$	$a_{**}(i)$

Table i	$Z=1$	$Z=0$	Totals
$Y=1$	$e_{11}(i)$	$e_{12}(i)$	$e_{1*}(i)$
$Y=0$	$e_{21}(i)$	$e_{22}(i)$	$e_{2*}(i)$
Totals	$e_{*1}(i)$	$e_{*2}(i)$	$e_{**}(i)$

Table 6.10: 2×2 table for ith age group of cases and controls. Left-hand table: observed counts. Right hand table: expected counts.

binomial sampling, Breslow shows that the variances of the individual contributions to the estimating equation are such that the quantity $2a_{**}^2(i)\text{Var}\,(R_i - \psi S_i)$ can be equated to;

$$E\left\{[a_{11}(i)a_{22}(i)+\psi a_{12}(i)a_{21}(i)]\,[a_{11}(i)+a_{22}(i)+\psi\,(a_{12}(i)+a_{21}(i))]\right\},$$

from which, by a simple application of the delta method we can obtain estimates of the variance of $\hat{\psi}_{MH}$.

Logistic regression

Without any loss in generality we can express the two probabilities of interest, $P(Y=1|Z=1)$ and $P(Y=1|Z=0)$ as simple power transforms of one another. This follows, since, whatever the true values of these probabilities, there exists some positive number α such that $P(Y=1|Z=1) = P(Y=1|Z=0)^\alpha$. The parameter α is constrained to be positive in order that the probabilities themselves remain between 0 and 1. To eliminate any potential dangers that may arise, particularly in the estimation context where, even though the true value of α is positive, the estimate itself may not be, a good strategy is to re-express this parameter as $\alpha = \exp(\beta)$. We then have

$$\log\log P(Y=1|Z=1) = \log\log P(Y=1|Z=0) + \beta. \qquad (6.31)$$

The parameter β can then be interpreted as a linear shift in the log-log transformation of the probabilities, and can take any value between $-\infty$ and ∞, the inverse transformations being one-to-one and guaranteed to lie in the interval (0,1). An alternative model to the above is

$$\text{logit}\,P(Y=1|Z=1) = \text{logit}\,P(Y=1|Z=0) + \beta. \qquad (6.32)$$

where the logit transformation, again one-to-one, is defined by $\text{logit}\,\theta = \log\{\theta/(1-\theta)\}$. Although a natural model, the model of Equation 6.31 is not usually preferred to that of Equation 6.32, motivated in an analogous way (i.e., avoiding constraints) but having a slight advantage from the viewpoint of interpretation. This is because the parameter β is the logarithm of the odds ratio, i.e., $\beta = \log \psi$.

In the light of the equivalence of the the odds for disease given the risk factor and the odds for the risk factor given the disease, as expressed in Equation 6.30, we conclude immediately that, equivalent to the above model involving β, expressed in Equation 6.32, we have a model expressing the conditional probability of Z given Y and using the same β. This highlights an important feature of proportional hazards modeling whereby we focus attention on the conditional distribution of the covariates given an event yet, when thinking of the applied physical problem behind the analysis, we would think more naturally in terms of the conditional distribution of the event given the covariates. The essential point is that the unknown regression parameter, β, of interest to us is the same for either situation so that in place of Equation (6.32), we can write

$$\operatorname{logit} P(Z = 1|Y = 1) = \operatorname{logit} P(Z = 1|Y = 0) + \beta. \tag{6.33}$$

Since the groups are indicated by a binary Z, we can exploit this in order to obtain the more concise notation, now common for such models, whereby

$$\operatorname{logit} P(Y = 1|Z) = \operatorname{logit} P(Y = 1|Z = 0) + \beta Z. \tag{6.34}$$

As we have tried, in as much as is possible throughout this text, to restrict attention to a single explanatory variable, this is once more the case here. Extension to multiple explanatory variables, or risk factors, is immediate and, apart from the notation becoming more cumbersome, there are no other concepts to which to give thought. We write the model down, as above in Equation (6.34), and use several binary factors Z (Z now a vector) to describe the different group levels. The coefficients β (β now a vector) then allow the overall odds ratio to be modeled or, allows the modeling of partial odds ratios whereby certain risk factors are included in the model, and our interest focuses on those remaining after having taken account of those already included. The above model can also be written in the form

$$\frac{P(Y = 1|Z)}{1 - P(Y = 1|Z)} = \exp(\beta_0 + \beta Z)\,, \tag{6.35}$$

where $\beta_0 = \operatorname{logit} P(Y = 1|Z = 0)$. Maintaining an analogy with the usual linear model we can interpret β_0 as an intercept, simply a function of the risk for a "baseline" group defined by $Z = 0$.

Assigning the value $Z = 0$ to some group and thereby giving that group baseline status is, naturally, quite arbitrary and there is nothing special about the baseline group apart from the fact that we define it as such. We are at liberty to make other choices and, in all events, the only quantities of real interest to us are relative ones. In giving thought to the different modeling possibilities that arise when dealing with a multivariate Z, the exact same kind of considerations, already described via several tables in the section on modeling multivariate problems will guide us (see Section 6.6 and those immediately following it). Rather than repeat or reformulate those ideas again here, the reader, interested in these aspects of epidemiological modeling, is advised to go over those earlier sections. Indeed, without a solid understanding as to why we choose to work with a particular model rather than another, and as to what the different models imply concerning the complex inter-relationships between the underlying probabilities, it is not really possible to carry out successful modeling in epidemiology.

Stratified and conditional logistic regression

In the above model, and Z being multivariate, we may wish to include alongside the main factors under study, known risk factors, and particularly risk factors such as age, or period effects, for which we would like to control. Often age alone is the strongest factor and its effect can be such that the associated errors of estimation in quantifying its impact can drown the effect of weaker risk factors. One possibility in controlling for such factors, S, it to appeal to the idea of stratification. This means that analysis is carried out at each level of S and, within a level, we make the same set of assumptions concerning the principle factors under study. We write

$$\frac{P(Y = 1|Z, S)}{1 - P(Y = 1|Z, S)} = \exp(\beta_0 + \beta Z)\,, \tag{6.36}$$

where, in the same way as before, $\beta_0 = \operatorname{logit} P(Y = 1|Z = 0, S)$. The important aspect of a stratified model is that the levels of S only appear in the left-hand side of the equation.

We might conclude that this is the same model as the previous one but it is not quite and, in later discussions on inference, we see that it does impact the way the likelihood is written. In the simpler cases, in as far as β is concerned, the stratified model is exactly equivalent to a regular logistic model if we include in the regression function indicator variables, of dimension one less than the number of strata. However, when the number of strata is large, use of the stratified model enables us to bypass estimation of the stratum-level effects. If these are not of real interest then this may be useful in that it can result in gains in estimating efficiency even though the underlying models may be equivalent. In a rough intuitive sense we are spending the available estimating power on the estimation of many less parameters, thereby increasing the precision of each one. This underlines an important point in that the question of stratification is more to do with inference than the setting up of the model itself.

This last remark is even more true when we speak of conditional logistic regression. The model will look almost the same as the unconditional one but the process of inference will be quite different. Suppose we have a large number of strata, very often in this context defined by age. A full model would be as in Equation (6.35), including in addition to the risk factor vector Z, a vector parameter of indicator variables of dimension one less than the number of strata. Within each age group, for the sake of argument let's say age group i, we have the simple logistic model. However, rather than write down the likelihood in terms of the products $P(Y=1|Z)$ and $P(Y=0|Z)$ we consider a different probability upon which to construct the likelihood, namely the probability that the event of interest, the outcome or case in other words, occurred on an individual (in particular the very individual for whom the event *did* occur, given that one event occurred among the set $S\{i\}$ of the $a_{**}(i)$ cases and controls. Denoting Z_i to be the risk factor for the case, corresponding to the age group i, then this probability is simply; $\exp(\beta Z_i)/\sum I[j \in S\{i\}]\exp(\beta Z_j)$. The likelihood is then the product of such terms across the number of different age groups for which a case was selected. If we carefully define the "at-risk" indicator $Y(t)$ where t now represents age, we can write the conditional likelihood as

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta Z_i)}{\sum_{j=1}^{n} Y_j(X_i)\exp(\beta Z_j)} \right\}^{\delta_i} . \tag{6.37}$$

Here we take the at-risk indicator function to be zero unless, for the subject j, X_j has the same age, or is among the same age group as that given by X_i. In this case the at-risk indicator $Y_j(X_i)$ takes the value one. To begin with, we assume that there is only a single case per age group, that the ages are distinct between age groups, and that, for individual i, the indicator δ_i takes the value one if this individual is a case. Use of the δ_i would enable us to include in an analysis sets of controls for which there was no case. This would be of no value in the simplest case but, generalizing the ideas along exactly the same lines as for standard proportional hazards models, we could easily work with indicators $Y(t)$ taking the value one for all values less than t and becoming zero if the subject becomes incident or is removed from the study. A subject is then able to make contributions to the likelihood at different values of t, i.e., at different ages, and appears therefore in different sets of controls. Indeed, the use of the risk indicator $Y(t)$ can be generalized readily to other complex situations.

One example is to allow it to depend on two time variables, for example, an age and a cohort effect, denoting this as $Y(t,u)$. Comparisons are then made between individuals having the same age and cohort status. Another useful generalization of $Y(t)$ is where individuals go on and off risk, either because they leave the risk set for a given period or, possibly, because their status cannot be ascertained. Judicious use of the at-risk indicator Y makes it possible then to analyze many types of data that, at first glance, would seem quite intractable. This can be of particular value in longitudinal studies involving time-dependent measurements where, in order to carry out unmodified analysis we would need, at each observed failure time, the time dependent covariate values for all subjects at risk. These would not typically all be available. A solution based on interpolation, assuming that measurements do not behave too erratically, is often employed. Alternatively we can allow for subjects for whom, at an event time, no reliable measurement is available, to simply temporarily leave the risk set, returning later when measurements have been made.

The striking thing to note about the above conditional likelihood is that it coincides with the expression for the partial likelihood given earlier in the chapter. This is no real coincidence of course and the main theorem of proportional hazards regression (Section 7.4), described in the following chapter, applies equally well here. For this we need one more concept, described later, and that is the idea of sampling from the risk set. The difference between the $Y(t)$ in a classical survival study, where it is equal to one as long as the subject is under study and then drops to zero, as opposed to the $Y(t)$ in the simple epidemiological application in which it is zero most of time, taking the value one when indicating the appropriate age group, is a small one. It can be equated with having taken a small random sample from a conceptually much larger group followed since time (age) is zero. On the basis of the above conditional likelihood we obtain the estimating equation

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left\{ Z_i - \frac{\sum_{j=1}^{n} Y_j(X_i) Z_j \exp(\beta Z_j)}{\sum_{j=1}^{n} Y_j(X_i) \exp(\beta Z_j)} \right\}, \tag{6.38}$$

which we equate to zero in order to estimate β. The equation contrasts the same quantities written down in Table 6.10 in which the expectations are taken with respect to the model. The estimating equations are then essentially the same as those given in Table 6.10 for the Mantel-Haenszel estimator. Furthermore, taking the second derivative of the expression for the log-likelihood, we have that $I(\beta) = \sum_{i=1}^{n} \delta_i I_i(\beta)$ where

$$I_i(\beta) = \frac{\sum_{j=1}^{n} Y_j(X_i) Z_j^2 \exp(\beta Z_j)}{\sum_{j=1}^{n} Y_j(X_i) \exp(\beta Z_j)} - \left\{ \frac{\sum_{j=1}^{n} Y_j(X_i) Z_j \exp(\beta Z_j)}{\sum_{j=1}^{n} Y_j(X_i) \exp(\beta Z_j)} \right\}^2, \tag{6.39}$$

then $I(\beta) = \sum_{i=1}^{n} \delta_i I_i(\beta)$. Inferences can then be carried out on the basis of these expressions. In fact, once we have established the link between the applied problem in epidemiology and its description via a proportional hazards model, we can then appeal to those model-building techniques (explained variation, explained randomness, goodness of fit, conditional survivorship function etc.) which we use for applications in time to event analysis. In this context the building of models in epidemiology is no less important, and no less delicate, than the building of models in clinical research.

6.12 Exercises and class projects

1. One of the early points of discussion on Cox's 1972 paper was how to deal with tied data. Look up the Cox paper and write down the various different ways that Cox and the contributors to the discussion suggested that tied data be handled. Explain the advantages and disadvantages to each approach.

2. One suggestion for dealing with tied data, not in that discussion, is to simply break the ties via some random split mechanism. What are the advantages and drawbacks to such an approach?

3. As an alternative to the proportional hazards model consider the two models (i) $S(t|Z) = S_0(t) + \beta Z$, and (ii) $\mathrm{logit} S(t|Z) = \mathrm{logit} S_0(t) + \beta Z$. Discuss the relative advantages and drawbacks of all three models.

4. Show that the relation; $S(t|Z) = \{S_0(t)\}^{\exp(\beta Z)}$ implies the Cox model and vice versa.

5. Suppose that we have two groups and that a proportional hazards model is believed to apply. Suppose also that we know for one of the groups that the hazard rate is a linear function of time, and equal to zero at the origin. Given data from such a situation, suggest different ways in which it can be analyzed and the possible advantages and disadvantages of the various approaches.

6. Explain in what sense the components of Equation 6.7 and equation (6.8) can be viewed as an equation for the mean and an equation for the variance.

7. Using equations (6.7) and (6.8) work out the calculations explicitly for the two-group case, i.e., the case in which there are $n_1(t)$ subjects at risk from group 1 at time t and $n_2(t)$ from group 2.

8. Suppose that we have available software able to analyze a proportional hazards model with a time-dependent covariate $Z(t)$. Suppose that, for the problem in hand the covariate, Z, does not depend on time. However, the regression effect $\beta(t)$ is known to decline as an exponential function of time. How would you proceed?

9. Suppose we fit a proportional hazards model, using some standard software, to a continuous covariate Z defined on the interval (1,4). Unknown to us our model assumption is incorrect and the model applies exactly to $\log Z$ instead. What effect does this have on our parameter estimate?

10. Consider an experiment in which there are eight levels of treatment. The levels are ordered. The null hypothesis is that there is no treatment effect. The alternative is that there exists a non-null effect increasing with level until it reaches one of the levels, say level j, after which the remaining levels all have the same effect as level j. How would you test for this?

11. Write down the joint likelihood for the underlying hazard rate and the regression parameter β for the two-group case in which we assume the saturated piecewise exponential model. Use this likelihood

to recover the partial likelihood estimate for β. Obtain an estimate of the survivorship function for both groups.

12. For the previous question derive an approximate large sample confidence interval for the estimate of the survivorship function for both groups in cases: (i) where the parameter β is exactly known, (ii) where the parameter is replaced by an estimate with approximate large sample variance σ^2.

13. Carry out a large sample simulation for a model with two binary variables. Each study is balanced with a total of 100 subjects. Choose $\beta_1 = \beta_2 = 1.5$ and simulate binary Z_1 and Z_2 to be uncorrelated. Show the distribution of $\hat{\beta}_1$ in two cases: (i) where the model used includes Z_2, (ii) where the model used includes only Z_1. Comment on the distributions, in particular the mean value of $\hat{\beta}_1$ in either case.
14. In the previous exercise, rather than include in the model Z_2, use Z_2 as a variable of stratification. Repeat the simulation in this case for the stratified model. Comment on your findings.

15. Consider the following regression situation. We have one-dimensional covariates Z, sampled from a density $g(z)$. Given z we have a proportional hazards model for the hazard rates. Suppose that, in addition, we are in a position to know exactly the marginal survivorship function $S(t) = \int S(t|z)g(z)dz$. How can we use this information to obtain a more precise analysis of data generated under the PH model with Z randomly sampled from $g(z)$?

16. Suppose we have two groups defined by the indicator variable $Z = \{0, 1\}$. In this example, unlike the previous in which we know the marginal survival, we know the survivorship function $S_0(t)$ for one of the groups. How can this information be incorporated into a two-group comparison in which survival for both groups is described by a proportional hazards model? Use a likelihood approach.

17. Use known results for the exponential regression model in order to construct an alternative analysis to that of the previous question based upon likelihood.

18. A simple test in the two-group case for absence of effects is to calculate the area between the two empirical survival curves. We can

evaluate the null distribution by permuting the labels corresponding to the group assignment indicator Z. Carry out this analysis for the Freireich data and obtain a p-value. How does this compare with that obtained from an analysis based on the assumption of an exponential model and that based on partial likelihood?

19. Carry out a study, i.e., the advantages, drawbacks and potentially restrictive assumptions, of the test of the previous example. How does this test compare with the score test based on the proportional hazards model?

20. Obtain a plot of the likelihood function for the Freireich data. Using simple numerical integration routines, standardize the area under the curve to be equal to one.

21. For the previous question, treat the curve as a density. Use the mean as an estimate of the unknown β. Use the upper and lower 2.5% percentiles as limits to a 95% confidence interval. Compare these results with those obtained using large sample theory.

22. Suppose we have six ordered treatment groups indicated by $Z = 1, \ldots, 6$. For all values of $Z \leq \ell$ the hazards are the same. For $Z > \ell$ the hazards are again the same and either the same as those for $Z \leq \ell$ or all strictly greater than for $Z \leq \ell$. The value of ℓ is not known. How would you model and set up tests in this situation?

23. Consider an epidemiological application in which workers may be exposed to some carcinogen during periods in which they work in some particular environment. When not working in that particular environment their risk falls back to the same as that for the reference population. Describe this situation via a proportional hazards model with time-dependent effects. How do you suggest modifying such a model if the risk from exposure rather than falling back to the reference group once exposure is removed is believed to be cumulative?

24. Write down a conditional logistic model in which we adjust for both age and cohort effects where cohorts are grouped by intervals of births from 1930-35, 1936-40, 1940-45, etc. For such a model is it possible to answer the question: was there a peak in relative risk during the nineteen sixties?