# Chapter 7

# Inference: Estimating equations

## 7.1 Summary

The results of this chapter and, for the most, all of the succeeding chapters, are based on an elementary and central theorem. We call this theorem the main theorem of proportional hazards regression. Its development is essentially that of O'Quigley (2003) which generalizes earlier results of Schoenfeld (1980), O'Quigley and Flandre (1994) and Xu and O'Quigley (2000). The theorem has several immediate corollaries and we can use these to write down estimating equations upon which we can then construct suitable inferential procedures for our models. While a particular choice of estimating equation can result in high efficiency when model assumptions are correct or close to being correct, other equations may be less efficient but still provide estimates which can be interpreted when model assumptions are incorrect. For example, when the regression function $\beta(t)$ might vary with time we are able to construct an estimating equation, the solution of which provides an estimate of $\beta$, in the case where $\beta(t)$ is a constant $\beta$, and $E\{\beta(T)\}$, the average effect, in the case where $\beta(t)$ changes through time. It is worth underlining that the usual partial likelihood estimate fails to achieve this.

## 7.2 Motivation

The earlier chapter on marginal survival is important in its own right and we lean on the results of that chapter throughout this work. We need keep in mind the idea of marginal survival for two reasons: (1) it provides a natural backdrop to the ideas of conditional survival and (2), together with the conditional distribution of the covariate given $T = t$, we are able to consider the joint distribution of covariate and survival time $T$. Conditional survival, where we investigate the conditional distribution of survival given different potential covariate configurations, as well as possibly time elapsed, is a central concern. More generally we are interested in survival distributions corresponding to transitions from one state to another, conditional on being in some particular state or of having mapped out some particular covariate path. The machinery that will enable us to obtain insight into these conditional distributions is that of proportional hazards regression.

When we consider any data at hand as having arisen from some experiment the most common framework for characterizing the joint distribution of the covariate $Z$ and survival $T$ is one where the distribution of $Z$ is fixed and known, and the conditional survivorship distribution the subject of our inferential endeavors. In fact, as underlined in the main theorem of proportional hazards regression, just below, it is more useful to characterize the joint distribution of $Z$ and $T$ via the conditional distribution of $Z$ given $T = t$ and the marginal distribution of $T$. This is one of the reasons why, in the previous chapter, we dealt with the marginal distribution of $T$. We can construct estimating equations based on these ideas and from these build simple tests or make more general inferences.

One of the most intriguing aspects of the Cox model concerns estimation of the regression parameter $\beta$ while ignoring any precise specification of $\lambda_0(t)$. Otherwise, under a conditional independent censoring mechanism and a specified functional form for the underlying hazard $\lambda_0(t)$, likelihood methods, at least in principle, are straightforward. But mostly we prefer to relax assumptions concerning $\lambda_0(t)$, possibly considering it to be entirely unknown, and construct inference for $\beta$ that remains invariant to any change in $\lambda_0(t)$. Any such changes can be made to correspond to monotonic increasing transformations on $T$, in which case we can take inference procedures to be rank invariant. This follows since monotonic increasing transformations on the observed times $X_i$ will not affect the rank ordering.

## 7.3 The observations

Our data will consist of the observations $(Z_i(t), Y_i(t), (t \leq X_i), X_i\,;\, i = 1 \ldots n)$. The $Z_i$ are the covariates (possibly time dependent), the $X_i = \min(T_i, C_i)$, the observed survival which is the smallest of the censoring time and the actual survival time and the $Y_i(t)$ are time-dependent indicators taking the value one as long as the $i$th subject is at risk at time $t$ and zero otherwise. For the sake of large sample constructions we make $Y_i(t)$ to be left continuous. At some level we will be making an assumption of independence, an assumption that can be challenged via the data themselves, but that is often left unchallenged, the physical context providing the main guide. Mostly, we think of independence as existing across the indices $i\,(i = 1, \ldots, n)$, i.e., the triplets $\{Z_i(t), Y_i(t), X_i\,; i = 1, \ldots, n\}$. It is helpful to our notational construction to have:

**Definition 7.1** *Let $\mathcal{Z}(t)$ be a data-based step function of $t$, everywhere equal to zero except at the points $X_i$, $i = 1, ..., n$, at which the function takes the value $Z_i(X_i)$. We assume that $|Z_i|$ is bounded, if not the definition is readily broadened.*

The reason for this definition is to unify notation. Our practical interest will be on sums of quantities such as $Z_i(X_i)$ with $i$ ranging from 1 to $n$. Using the Stieltjes integral, we will be able to write such sums as integrals with respect to an empirical process. In view of the Helly-Bray theorem (Section 2.3) this makes it easier to gain an intuitive grasp on the population structure behind the various statistics of interest. Both $T$ and $C$ are assumed to have supports on some finite interval, the first of which is denoted $\mathcal{T}$. The time-dependent covariate $Z(\cdot)$ is assumed to be a left continuous stochastic process and, for notational simplicity, is taken to be of dimension one whenever possible. Let $F(t) = \Pr(T < t)$, $D(t) = \Pr(C < t)$ and $H(t) = F(t)\{1 - D(t)\} - \int_0^t F(u)dD(u)$.

For each subject $i$ we observe $X_i = \min(T_i, C_i)$, and $\delta_i = I(T_i \leq C_i)$ so that $\delta_i$ takes the value one if the $i$th subject corresponds to a failure and is zero if the subject corresponds to a censored observation. A more general situation allows a subject to be dynamically censored in that he or she can move in and out of the risk set. To do this we define the "at-risk" indicator $Y_i(t)$ where $Y_i(t) = I(X_i \geq t)$. The events on the $i$th individual are counted by $N_i(t) = I\{T_i \leq t, T_i \leq C_i\}$ and $\bar{N}(t) = \sum_1^n N_i(t)$ counts the number of events before $t$. It is also helpful to be able to refer to the total number of observed

failures $k = \bar{N}\{\sup t : t \in \mathcal{T}\}$, and the inverse function $\bar{N}^{-1}(\cdot)$, where $\bar{N}^{-1}(\ell) = \{\inf t : t \in \mathcal{T}, \bar{N}(t) = \ell\}$, the smallest time by which a given number of events $\ell$ have occurred. Consistent estimators of $F(t)$ and $H(t)$ are indicated by hats, the examples here being the Kaplan-Meier estimator for $1 - F(t)$ and $\hat{H}(t) = n^{-1}\bar{N}(t)$.

Some other sums of observations will frequently occur. In order to obtain an angle on empirical moments under the model, Andersen and Gill (1982) define

$$S^{(r)}(\beta, t) = n^{-1} \sum_{i=1}^{n} Y_i(t) e^{\beta Z_i(t)} Z_i(t)^r, \quad s^{(r)}(\beta, t) = E S^{(r)}(\beta, t),$$

for $r = 0, 1, 2$, where the expectations are taken with respect to the true distribution of $(T, C, Z(\cdot))$. Define also

$$V(\beta, t) = \frac{S^{(2)}(\beta, t)}{S^{(0)}(\beta, t)} - \frac{S^{(1)}(\beta, t)^2}{S^{(0)}(\beta, t)^2}, \; v(\beta, t) = \frac{s^{(2)}(\beta, t)}{s^{(0)}(\beta, t)} - \frac{s^{(1)}(\beta, t)^2}{s^{(0)}(\beta, t)^2}. \quad (7.1)$$

The Andersen and Gill notation is now classic in this context. Their notation lends itself more readily to large sample theory based upon martingales and stochastic integrals. We will frequently keep this notation in mind although our approaches to inference do not appeal to special central limit theorems (the martingale central limit theorem in particular) and, as a result, our notation is typically lighter. The required conditions for the Andersen and Gill theory to apply are sightly broader although this advantage is more of a theoretical than a practical one. For their results, as well as ours, the censorship is restricted in such a way that, for large samples, there remains information on $F$ in the tails. The conditional means and the conditional variances, $\mathcal{E}_{\beta(t)}(Z|t)$ $\mathcal{V}_{\beta(t)}(Z|t)$, introduced immediately below, are related to the above via $V(\beta, t) \equiv \mathcal{V}_{\beta}(Z|t)$ and $S^{(1)}(\beta, t)/S^{(0)}(\beta, t) \equiv \mathcal{E}_{\beta}(Z|t)$. In the counting process framework of Andersen and Gill (1982), we imagine $n$ as remaining fixed and the asymptotic results obtaining as a result of asymptotic theory for $n$-dimensional counting processes, in which we understand the expectation operator $E$ to be with respect to infinitely many repetitions of the process. Subsequently we allow $n$ to increase without bound. For the quantities $\mathcal{E}_{\beta(t)}(Z^k|t)$ we take the $E$ operator to be these same quantities when $n$ becomes infinitely large.

## 7.4 Main theorem

A simple theorem underpins all of the key results discussed in this book (testing the presence of regression effect, estimating average regression effect under non-proportional hazards, quantifying predictability via the conditional survivorship function as well as via summary indices such as explained randomness and explained variation, assessing fit, contrasting competing models etc). In view of all these several applications the theorem then appears to be quite fundamental and, as such, it seems appropriate to refer to it as the main theorem of proportional hazards regression.

We most often view time as providing the set of indices to certain stochastic processes, so that, for example, we consider $Z(t)$ to be a random variable having different distributions for different $t$. Also, the failure time variable $T$ can be viewed as a non-negative random variable with distribution $F(t)$ and, whenever the set of indices $t$ to the stochastic process coincide with the support for $T$, then not only can we talk about the random variables $Z(t)$ for which the distribution corresponds to $P(Z \leq z|T = t)$ but also marginal quantities such as the random variable $Z(T)$ having distribution $G(z) = P(Z \leq z)$. An important result concerning the conditional distribution of $Z(t)$ given $T = t$ follows. First we need the following definitions:

**Definition 7.2** *The discrete probabilities $\pi_i(\beta(t), t)$ are given by*

$$\pi_i(\beta(t), t) = \frac{Y_i(t) \exp\{\beta(t) Z_i(t)\}}{\sum_{j=1}^{n} Y_j(t) \exp\{\beta(t) Z_j(t)\}}. \tag{7.2}$$

The $\pi_i(\beta(t), t)$ are easily seen to be bona fide probabilities (for all real values of $\beta(t)$) since $\pi_i \geq 0$ and $\sum_i \pi_i = 1$. Note that this continues to hold for values of $\beta(t)$ different to those generating the data, and even when the model is incorrectly specified. As a consequence, replacing $\beta$ by $\hat{\beta}$ results in a probability distribution that is still valid but different to the true one. Means and variances with respect to this distribution maintain their interpretation as means and variances.

Under the proportional hazards assumption, i.e., the constraint $\beta(t) = \beta$, the product of the $\pi$'s over the observed failure times gives the partial likelihood (Cox 1972, 1975). When $\beta = 0$, $\pi_i(0, t)$ is the empirical distribution that assigns equal weight to each sample subject in the risk set. Based on the $\pi_i(\beta(t), t)$ we have:

**Definition 7.3** *Conditional moments of $Z$ with respect to $\pi_i(\beta(t), t)$ are given by*

$$\mathcal{E}_{\beta(t)}(Z^k|t) = \sum_{i=1}^{n} Z_i^k(t)\pi_i(\beta(t), t)\,,\ k = 1, 2, \ldots, . \tag{7.3}$$

These two definitions are all that we need in order to set about building the structures upon which inference is based. This is particularly so when we are able to assume an independent censoring mechanism, although the weaker assumption of a conditionally independent censoring mechanism (see Chapter 4) will mostly cause no conceptual difficulties; simply a slightly more burdensome notation. Another, somewhat natural, definition will also be appealed to on occasion and this concerns unconditional expectations.

**Definition 7.4** *Marginal moments of $Z$ with respect to the bivariate distribution characterized by $\pi_i(\beta(t), t)$ and $F(t)$ are given by*

$$\mathcal{E}_{\beta(t)}(Z^k) = \int \mathcal{E}_{\beta(t)}(Z^k|t)dF(t)\,,\ k = 1, 2, \ldots, . \tag{7.4}$$

Recall that for arbitrary random variables $A$ and $B$, assuming expectation to be defined, we have the result of double expectation whereby $E(A) = EE(A|B)$. This is the motivation behind the above definition. Once again, these expectations are to be interpreted as population quantities in as much as $\beta(t)$ and $F(t)$ are taken to be known. They can also, of course, be viewed as sample-based quantities since $n$ is finite and the $Y_i(t)$ are random until time point $t$. At the end of the study the paths of all the $Y_i(t)$ are known and we are, to use a common expression, "conditioning on the data." The art of inference, and its understanding, stem, to a great extent, from knowing which aspects of an experiment to view as random (given that once the experiment is over there is not really anything truly random). Also which distributions are relevant and these can change so that, here for example, we should think carefully about the meaning of the expectation operators $E$ and $\mathcal{E}$ in its particular context. These expectations are still well defined, but with respect to different distributions; when replacing $\beta$ by $\hat{\beta}$, when replacing $F$ by $F_n$ and $\hat{F}$, and when allowing $n$ to go to infinity. The quantity $\phi$ of the following definition is not of any essential interest, featuring in the main theorem but disappearing afterwards.

**Definition 7.5** *In order to distinguish conditionally independent censoring from independent censoring we define* $\phi(z,t)$ *where*

$$\phi(z^*,t) = \frac{\int P(C \geq t|z)g(z)dz}{P(C \geq t|z^*)}.$$

Note that when censoring does not depend upon $z$ then $\phi(z,t)$ will depend upon neither $z$ nor $t$ and is, in fact, equal to one. Otherwise, under a conditionally independent censoring assumption, we can consistently estimate $\phi(z,t)$ and we call this $\hat{\phi}(z,t)$. The following theorem is presented in O'Quigley (2003).

**Theorem 7.1** *Under model (6.2) and assuming* $\beta(t)$ *known, the conditional distribution function of* $Z(t)$ *given* $T = t$ *is consistently estimated by*

$$\hat{P}\{Z(t) \leq z|T=t\} = \frac{\sum_{z_i \leq z} Y_i(t)\exp\{\beta(t)z_i(t)\}\hat{\phi}(z_i,t)}{\sum_{j=1}^{n} Y_j(t)\exp\{\beta(t)z_j(t)\}\hat{\phi}(z_j,t)}. \qquad (7.5)$$

The theorem, which we refer to as the main theorem of proportional hazards regression, has many important consequences including:

**Corollary 7.1** *Under model (6.2) and an independent censorship, assuming* $\beta(t)$ *known, the conditional distribution function of* $Z(t)$ *given* $T = t$ *is consistently estimated by*

$$\hat{P}(Z(t) \leq z|T=t) = \sum_{j=1}^{n} \pi_j(\beta(t),t)I(Z_j(t) \leq z). \qquad (7.6)$$

The observation we would like to make here is that we can fully describe a random variable indexed by $t$, i.e., a stochastic process. All of our inference will follow from this. In essence, we first fix $t$ and then we fix our attention on the conditional distribution of $Z$ given that $T = t$ and models which enable us to characterize this distribution. Indeed, under the broader censoring definition of conditional independence, common in the survival context, we can still make the same basic observation. In this case we condition upon something more complex that just $T = t$ but the actual random outcome that we condition upon is of less importance than the simple fact that we are able to described sets of conditional distributions all indexed by $t$, i.e., a stochastic process indexed by $t$. Specifically:

**Corollary 7.2** *For a conditionally independent censoring mechanism we have*

$$\hat{P}(Z(t) \leq z | T = t, C > t) = \sum_{j=1}^{n} \pi_j(\beta(t), t) I(Z_j(t) \leq z). \qquad (7.7)$$

Whether we condition on the event $T = t$ or the event $(T = t, C > t)$, we identify a random variable indexed by $t$. This is all we need to construct appropriate stochastic processes (functions of $Z(t)$) enabling inference. Again simple applications of Slutsky's theorem shows that the result still holds for $\beta(t)$ replaced by any consistent estimate. In particular, when the hypothesis of proportionality of risks is correct, the result holds for the estimate $\hat{\beta}$. The following two corollaries follow immediately from those just above and form the basis to the main tests we construct. For integer $k$ we have:

**Corollary 7.3** $\mathcal{E}_{\hat{\beta}(t)}(Z^k|t)$ *provides a consistent estimate of* $E_{\beta(t)}(Z^k(t)|t)$*, under model (6.2). In particular* $\mathcal{E}_{\hat{\beta}}(Z^k|t)$ *provides a consistent estimate of* $E_{\beta}(Z^k(t)|t)$*, under the model expressed by Equation 6.3.*

Furthermore, once again working under the model, we consider:

**Definition 7.6** $\mathcal{V}_{\beta(t)}(Z|t) = \mathcal{E}_{\beta(t)}(Z^2|t) - \mathcal{E}^2_{\beta(t)}(Z|t)\,.$

In practical data analysis the quantity $\beta(t)$ may be replaced by a value constrained by some hypothesis or an estimate. The quantity $\mathcal{V}_{\beta(t)}(Z|t)$ can be viewed as a conditional variance which may vary little with $t$, in a way analogously to the residual variance in linear regression which, under classic assumptions, remains constant with different levels of the independent variable. Since $\mathcal{V}_{\beta(t)}(Z|t)$ may change with $t$, even if not a lot, it is of interest to consider some average quantity and so we also introduce:

**Definition 7.7** $E\,\mathcal{V}_{\beta(t)}(Z) = \int \mathcal{V}_{\beta(t)}(Z|t) dF(t)\,.$

These sample-based variances relate to population variances via the following corollary;

**Corollary 7.4** *Under model (6.3),* $\mathrm{Var}(Z|t)$ *is consistently estimated by* $\mathcal{V}_{\hat{\beta}}(Z|t)$*.* $E\,\mathrm{Var}(Z|t)$ *is consistently estimated by* $E\,\mathcal{V}_{\hat{\beta}}(Z|t)$*. In addition,* $\int \mathcal{V}_{\hat{\beta}}(Z|t) d\hat{F}(t)$ *is consistent for* $E\,\mathrm{Var}(Z|t)$*.*

These quantities are all useful in our construction. Interpretation requires some care. For example, although $E\,\mathcal{V}_{\hat{\beta}}(Z|t)$ is, in some sense, a marginal quantity, it is not the marginal variance of $Z$ since we have neglected the variance of $E_{\beta(t)}(Z(t)|t)$ with respect to the distribution of $T$. The easiest case to interpret is the one where we have an independent censoring mechanism (Equation 7.6). However, we do not need to be very concerned about any interpretation difficulty, arising for instance in Equation 7.7 where the censoring time appears in the expression, since, in this or the simpler case, all that matters to us is that our observations can be considered as arising from some process, indexed by $t$ and, for this process, we are able, under, as usual, some model assumptions, to consistently estimate the mean and the variance of the quantities that we observe. It is also useful to note another natural relation between $\mathcal{V}_\beta(Z|t)$ and $\mathcal{E}_\beta(Z|t)$ since

$$\mathcal{V}_\beta(Z|t) = \partial\,\mathcal{E}_\beta(Z|t)/\partial\beta.$$

This relation is readily verified for fixed $\beta$. In the case of time-dependent $\beta(t)$ then, at each given value of $t$, it is again clear that the same relation holds. The result constitutes one of the building blocks in the overall inferential construction and, under weak conditions, for example $Z$ being bounded, then it also follows that

$$\int \mathcal{V}_\beta(Z|t) = \int \partial\,\mathcal{E}_\beta(Z|t)/\partial\beta = \partial\left\{\int\,\mathcal{E}_\beta(Z|t)\right\}/\partial\beta.$$

Throughout the rest of this book we will see just why the main theorem is so fundamental. Essentially all the information we need, for almost any conceivable statistical goal, arising from considerations of any of the models considered, is contained in the joint probabilities $\pi_i(\beta(t), t)$ of the fundamental definition 7.2. We are often interested, in the multivariate setting for example, in the evaluation of the effects of some factor while having controlled for others. This can be immediately accommodated. Specifically, taking $Z$ to be of some dimension greater than one ($\beta$ being of the same dimension) and writing $Z^T = (Z_1^T, Z_2^T)$ and $Z_i^T = (Z_{1i}^T, Z_{2i}^T)$ then, summing over the multivariate probabilities, we have two obvious extensions to Corollaries 7.1 and 7.2.

**Corollary 7.5** *Under model (6.2) and an independent censorship, assuming $\beta(t)$ known, the conditional distribution function of $Z_2(t)$ given $T = t$ is consistently estimated by*

$$\hat{P}(Z_2(t) \leq z|T = t) = \sum_{j=1}^{n} \pi_j(\beta(t), t) I(Z_{2j}(t) \leq z). \tag{7.8}$$

The corollary enables component wise inference. We can consider the components of the vector $Z_i$ individually. Also we could study some functions of the components, usually say a simple linear combination of the components such as the prognostic index. Note also that:

**Corollary 7.6** *For a conditionally independent censoring mechanism we have*

$$\hat{P}(Z_2(t) \leq z|T = t, C > t) = \sum_{j=1}^{n} \pi_j(\beta(t), t) I(Z_{2j}(t) \leq z), \tag{7.9}$$

where in Definition 7.2 for $\pi_j(\beta(t), t)$ we take $\beta(t)Z_j(t)$ to be an inner product, which we may prefer to write as $\beta(t)^T Z_j(t)$ and where $Z_j(t)$ are the observed values of the vector $Z(t)$ for the $j$th subject. Also, by $Z_2(t) \leq z$ we mean that all of the scalar components of $Z_2(t)$ are less than or equal to the corresponding scalar components of $z$. As for the corollaries and definitions following Corollaries 7.1 and 7.2 they have obvious equivalents in the multivariate setting and so we can readily write down expressions for expectations, variances and covariances as well as their corresponding estimates.

*Moments for stratified models*

Firstly we recall from the previous chapter that the stratified model is simply a partially proportional hazards model in which some of the components of $\beta(t)$ remain unspecified while the other components are constant terms. The definition for the stratified model was

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

where $s$ takes integer values 1,..., $m$. In view of the equivalence between stratified models and partially proportional hazards models described in the previous chapter, the main theorem and its corollaries apply immediately. However, in light of the special importance of stratified models, as proportional hazards models with relaxed assumptions, it will be helpful to our development to devote a few words to this case. Analogous to the above definition for $\pi_i(\beta(t), t)$, and using the, possibly time-dependent, stratum indicator $s(t)$ we now define these probabilities via:

**Definition 7.8** *For the stratified model, having strata* $s = 1, \ldots, m$ *the discrete probabilities* $\pi_i(\beta(t), t)$ *are now given by*

$$\pi_i(\beta(t), t) = \frac{Y_i\{s(t), t\} \exp\{\beta(t) Z_i(t)\}}{\sum_{j=1}^{n} Y_j\{s(t), t\} \exp\{\beta(t) Z_j(t)\}}. \tag{7.10}$$

When there is a single stratum then this definition coincides with the earlier one and, indeed, we use the same $\pi_i(\beta(t), t)$ for both situations, since it is only used indirectly and there is no risk of confusion. Under equation (6.3), i.e. the constraint $\beta(t) = \beta$, the product of the $\pi$'s over the observed failure times gives the so-called stratified partial likelihood (Kalbfleisch and Prentice 1980). The series of above definitions for the non-stratified model, in particular Definition 7.2, theorems, corollaries, all carry over in an obvious way to the stratified model and we do not propose any additional notation. It is usually clear from the context although it is worth making some remarks. Firstly, we have no direct interest in the distribution of $Z$ given $t$ (note that this distribution depends on the distribution of $Z$ given $T > 0$, a distribution which corresponds to our design and is quite arbitrary).

We will exploit the main theorem in order to make inferences on $\beta$ and, in the stratified case, we would also condition upon the strata from which transitions can be made. In practice, we contrast the observations $Z_i(X_i)$, made at time point $X_i$ at which an event occurs ($\delta_i = 1$) with those subjects at risk of the same event. The "at risk" indicator, $Y(s(t), t)$, makes this very simple to express. We can use $Y(s(t), t)$ to single out appropriate groups for comparison. This formalizes a standard technique in epidemiology whereby the groups for comparison may be matched by not just age but by other variables. Such variables have then been controlled for and eliminated from the analysis. Their own specific effects can be quite general and we are not in a position to estimate them. Apparently very complex situations, such as subjects moving in and out of risk categories, can be easily modeled by the use of these indicator variables.

*Moments for other relative risk models*

Instead of Equation 6.2 some authors have suggested a more general form for the hazard function whereby

$$\lambda(t|Z) = \lambda_0(t) R\{\beta(t) Z\}, \tag{7.11}$$

and where, mostly, $\beta(t)$ is not time-varying, being equal to some unknown constant. The most common choices for the function $R(r)$ are $\exp(r)$, in which case we recover the usual model, and $1+r$ which leads to the so-called additive model. Since both $\lambda(t|Z)$ and $\lambda_0$ are necessarily positive we would generally need constraints on the function $R(r)$. In practice this can be a little bothersome and is, among several other good reasons, a cause for favoring the multiplicative risk model $\exp(r)$ over the additive risk model $1 + r$. If we replace our earlier definition for $\pi_i(\beta(t), t)$ by:

**Definition 7.9** *The discrete probabilities $\pi_i(\beta(t), t)$ are given by;*

$$\pi_i(\beta(t), t) = \frac{Y_i(t)R\{\beta(t)Z_i(t)\}}{\sum_{j=1}^{n} Y_j(t)R\{\beta(t)Z_j(t)\}}, \tag{7.12}$$

then all of the above definitions, theorems, and corollaries have immediate analogues and we do not write them out explicitly. Apart from one interesting exception, which we look at more closely in the chapters dealing with inference, there are no particular considerations we need concern ourselves over if we choose $R(r) = 1 + r$ rather than $R(r) = \exp(r)$. Note also that if we allow the regression functions, $\beta(t)$, to depend arbitrarily upon time then, given either model, the other model exists with a different function of $\beta(t)$. The only real reason for preferring one model over another would be due to parsimony; for example, we might find in some given situation that in the case of the additive model the regression function $\beta(t)$ is in fact constant unlike the multiplicative model where it may depend on time. But otherwise both functions may depend, at least to some extent, on time and then the multiplicative model ought be preferred since it is the more natural. We say the more natural because the positivity constraint is automatically satisfied.

*Transformed covariate models*

For some transformation $\psi$ of the covariate we can postulate a model of the form;

$$\lambda(t|Z(t)) = \lambda_0(t) \exp\{\beta(t)\psi[Z(t)]\}. \tag{7.13}$$

All of the calculations proceed as above and no real new concept is involved. Such models can be considered in the case of continuous covariates, $Z$, which may be sufficiently asymmetric, implying very great

changes of risk at the high or low values, to be unlikely to provide a satisfactory fit. Taking logarithms, or curbing the more extreme values via a defined plateau, or some other such transformation will produce models of potentially wider applicability. Note that this is a different approach to working with, say,

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

and using the main theorem, in conjunction with estimating equations described here below and basing inference upon the observations $\psi Z(X_i)$ and their expectations under this model. In this latter case we employ $\psi$ in the estimating equation as a means to obtain greater robustness or to reduce sensitivity to large observations. In the former case the model itself is different and would lead to different estimates of survival probabilities.

Our discussion so far has turned around the hazard function. However, it is equally straightforward to work with intensity functions and these allow for increased generality, especially when tackling complex time-dependent effects. O'Brien (1978) introduced the logit-rank test for survival data when investigating the effect of a continuous covariate on survival time. His purpose was to construct a test that was rank invariant with respect to both time and the covariate itself. O'Quigley and Prentice (1991) showed how a broad class of rank invariant procedures can be developed within the framework of proportional hazards models. The O'Brien logit-rank procedure was a special case of this class. In these cases we work with intensity rather than hazard functions. Suppose then that $\lambda_i(t)$ indicates an intensity function for the $i$th subject at time $t$. A proportional hazards model for this intensity function can be written

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{\beta Z_i(t)\},$$

where $Y_i(t)$ indicates whether or not the $i^{th}$ subject is at risk at time $t$, $\lambda_0(t)$ the usual "baseline" hazard function and $Z_i(t)$ is a constructed covariate for the $i$th subject at time $t$. Typically, $Z_i(t)$ in the estimating equation is defined as a function of measurements on the $i$th subject alone, but it can be defined more generally as $Z_i(t) = \psi_i(t, \mathcal{F}_t)$ for $\psi$ some function of $\mathcal{F}_t$, the collective failure, censoring and covariate information prior to time $t$ on the entire study group. The examples in O'Quigley and Prentice (1991) included the rank of the the subject's covariate at $X_i$ and transformations on this such as the normal order

statistics. This represents a departure from most regression situations because the value used in the estimating equation depends not only on what has been observed on the particular individual but also upon what has been observed on other relevant subsets of individuals.

*Misspecified models*

For multinormal linear regression involving $p$ regressors we can eliminate from consideration some of these and focus our attention on models involving the remaining regressors strictly less than $p$. We could eliminate these by simple integration, thereby obtaining marginal distributions. Under the usual assumptions of multiple linear regression the resulting lower dimensional model remains a multinormal one. As an example, in the simple case of a two dimensional covariate normal model, both the marginal models involving only one of the two covariates are normal models. However, for non-linear models this result would only be expected to hold under quite unusual circumstances. Generally, for non-linear models, and specifically proportional hazards models, the result will not hold so that if the model is assumed true for a covariate vector of dimension $p$, then, for any submodel, of dimension less that $p$, the model will not hold exactly. A corollary to this is that no model of dimension greater that $p$ could exactly follow a proportional hazards prescription if we claim that the model holds precisely for some given $p$ covariates.

These observations led some authors to claim that "forgotten" or "overlooked" variables would inevitably lead to misleading results. Such a claim implies that *all* analyses based on proportional hazards models are misleading and since, to say the least, such a conclusion is unhelpful we offer a different perspective. This says that *all* practical models are only ever approximately correct. In other words, the model is always making a simplifying assumption, necessarily overlooking potential effects as well as including others which may impact the proportionality of those key variables of interest. Our task then focuses on interpreting our estimates when our model cannot be exactly true. In terms of analysing real data, it makes much more sense to take as our underlying working assumption that the model is, to a greater or lesser degree, misspecified.

A model can be misspecified in one of two clear ways; the first is that the covariate form is not correctly expressed and the second is that the regression coefficient is not constant through time. An ex-

ample of the first would be that the true model holds for $\log Z$ but that, not knowing this, we include $Z$ in the model. An example of the second might have $\beta(t)$ declining through time rather than remaining constant.

It has been argued that the careful use of residual techniques can indicate which kind of model failure may be present. This is not so. Whenever a poor fit could be due to either cause it is readily seen that a misspecified covariate form can be represented correctly via a time-dependent effect. In some sense the two kinds of misspecification are unidentifiable. We can fix the model by working either with the covariate form or the regression coefficient $\beta(t)$. Of course, in certain cases, a discrete binary covariate describing two groups, for example, there can only be one cause of model failure - the time dependency of the regression coefficient. This is because the binary coding imposes no restriction of itself since all possible codings are equivalent.

The important issue is then the interpretation of an estimate, say $\hat{\beta}$ under a proportional hazards assumption when, in reality, the data are generated under the broader non-proportional hazards model with regression coefficient function $\beta(t)$. This is not a straightforward endeavor and the great majority of the currently used procedures, including those proposed in the widely distributed R, SAS, STATA and S-Plus packages, produce estimates which cannot be interpreted unless there is no censoring. To study this question we first define $\mu = \int \beta(t)dF(t)$, which is an average of $\beta(T)$ with respect to the distribution $F(t)$. It is also of interest to consider the approximation

$$\hat{P}(Z(t) \leq z | T = t, C > t) \approx \sum_{j=1}^{n} \pi_j(\mu, t) I(Z_j(t) \leq z) \qquad (7.14)$$

and, for the case of a model making the stronger assumption of an independent censoring mechanism as opposed to a conditionally independent censoring mechanism given the covariate, we have

$$\hat{P}(Z(t) \leq z | T = t) \approx \sum_{j=1}^{n} \pi_j(\mu, t) I(Z_j(t) \leq z). \qquad (7.15)$$

For small samples it will be unrealistic to hope to obtain reliable estimates of $\beta(t)$ for all of $t$ so that, often, we take an estimate of some summary measure, in particular $\mu$. It is in fact possible to construct an estimating equation which provides an estimate of $\mu$ without estimating $\beta(t)$ (Xu and O'Quigley 1998) and it is very important to

stress that, unless there is no censoring, the usual estimating equation which leads to the partial likelihood estimate does not accomplish this. In fact, the partial likelihood estimate turns out to be equivalent to obtaining the solution of an estimating equation based on $H(t)$ (see Section 7.3) and using $\hat{H}(t)$ as an estimate whereas, to consistently estimate $\mu$, it is necessary to work with some consistent estimate of $F(t)$, in particular the Kaplan-Meier estimate.

Some thought needs be given to the issues arising when our estimating equation is based on certain assumptions (in particular, a proportional hazards assumption), whereas the data themselves can be considered to have been generated by something broader (in particular, a non proportional hazards model). To this purpose we firstly consider a definition that will allow us to anticipate just what is being estimated when the data are generated by model (6.2) and we are working with model (6.3). This is contained in the definition for $\beta^*$ just below.

Let's keep in mind the widely held belief that the partial likelihood estimate obtained when using a proportional hazards model in a situation where the data are generated by a broader model must correspond to some kind of average effect. It does correspond to something (as always) but nothing very useful and not something we can hopefully interpret as an average effect. This is considered in the following sections. Firstly we need:

**Definition 7.10** *Let $\beta^*$ be the constant value satisfying*

$$\int_{\mathcal{T}} \mathcal{E}_{\beta^*}(Z|t)dF(t) = \int_{\mathcal{T}} \mathcal{E}_{\beta(t)}(Z|t)dF(t). \tag{7.16}$$

The definition enables us to make sense out of using estimates based on (6.3) when the data are in fact generated by (6.2). Since we can view $T$ as being random, whenever $\beta(t)$ is not constant, we can think of having sampled from $\beta(T)$. The right-hand side of the above equation is then a double expectation and $\beta^*$, occurring in the left-hand side of the equation, is the best fitting value under the constraint that $\beta(t) = \beta$. We can show the existence and uniqueness of solutions to Equation (7.16) (Xu and O'Quigley 1998). More importantly, $\beta^*$ can be shown to have the following three properties: (i) under model (6.3) $\beta^* = \beta$; (ii) under a subclass of the broad class of models known as the Harrington-Fleming models, we have an exact result in that $\beta^* = \int_{\mathcal{T}} \beta(t)dF(t)$; and (iii) for very general situations we can write that

$\beta^* \approx \int_{\mathcal{T}} \beta(t) dF(t)$, an approximation which is in fact very accurate. Estimates of $\beta^*$ are discussed in (Xu and O'Quigley 1998, Xu and O'Quigley 2000) and, in the light of the foregoing, we can take these as estimates of $\mu$.

Theorem 7.1 and its corollaries provide the ingredients necessary to constructing a number of relevant stochastic processes, in particular functions of Brownian motion. We will be able to construct a process that will look like simple Brownian motion under the chosen model and with given parameter values. We can then consider what this process will look like when, instead of those null values, the data are generated by a model from the same class but with different parameter values. First we consider the estimating equations that can be readily constructed as a result of the preceding theory.

## 7.5 The estimating equations

The above setting helps us anticipate the properties of the estimators we will be using. First, recall our definition of $\mathcal{Z}(t)$ as a step function of $t$ with discontinuities at the points $X_i$, $i = 1, ..., n$, at which the function takes the value $Z_i(X_i)$. Next, consider $F_n(t)$, the empirical marginal distribution function of $T$. Note that $F_n(t)$ coincides with the Kaplan-Meier estimate of $F(t)$ in the absence of censoring. When there is no censoring, a sensible estimating equation (which we will see also arises as the derivative of a log likelihood, as well as the log partial likelihood) is

$$U_1(\beta) = \int \{\mathcal{Z}(t) - \mathcal{E}_\beta(Z|t)\} dF_n(t) = 0. \tag{7.17}$$

The above integral is simply the difference of two sums, the first the empirical mean without reference to any model and the second the average of model-based means. It makes intuitive sense as an estimating equation and the only reason for writing the sum in the less immediate form as an integral is that it helps understand the large sample theory when $F_n(t) \xrightarrow{\text{p}} F(t)$. Each component in the above sum includes the size of the increment, $1/n$, a quantity that can then be taken outside of the summation (or integral) as a constant factor. Since the right-hand side of the equation is identically equal to zero, the incremental size $1/n$ can be canceled, enabling us to rewrite the equation as

$$U_2(\beta) = \int \{\mathcal{Z}(t) - \mathcal{E}_\beta(Z|t)\} d\bar{N}(t) = 0. \tag{7.18}$$

It is this expression where the integral is taken with respect to increments $d\bar{N}(t)$, rather than with respect to $dF_n(t)$ that is the more classic representation in this context. The expression equates $U_2(\beta)$ in terms of the counting processes $N_i(t)$. These processes, unlike the empirical distribution function, are available in the presence of censoring. It is the above equation that is used to define the partial likelihood estimator, since, unless the censoring is completely absent, the quantity $U_1(\beta)$ is not defined.

A natural question would be the following: suppose two observers were to undertake an experiment to estimate $\beta$. A certain percentage of observations remain unobservable to the first observer as a result of an independent censoring mechanism but are available to the second observer. The first observer uses Equation 7.18 to estimate $\beta$, whereas the second observer uses Equation 7.17. Will the two estimates agree? By "agree" we mean, under large sample theory, will they converge to the same quantity. We might hope that they would; at least if we are to be able to usefully interpret estimates obtained from Equation 7.18. Unfortunately though (especially since Equation 7.18 is so widely used), the estimates do not typically agree. Table 7.1 below indicates just how severe the disagreement might be. However, the form of $U_1(\beta)$ remains very much of interest and, before discussing the properties of

Table 7.1: Comparison of $\beta^*$, $\int \beta(t)dF(t)$, and the estimates $\tilde{\beta}$ and $\hat{\beta}_{PL}$

| $\beta_1$ | $\beta_2$ | $t_0$ | % censored | $\beta^*$ | $\int \beta(t)dF(t)$ | $\tilde{\beta}$ | $\hat{\beta}_{\mathrm{PL}}$ |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 0.1 | 0% | 0.156 | 0.157 | 0.155 (0.089) | 0.155 (0.089) |
| | | | 17% | 0.156 | 0.157 | 0.158 (0.099) | 0.189 (0.099) |
| | | | 34% | 0.156 | 0.157 | 0.160 (0.111) | 0.239 (0.111) |
| | | | 50% | 0.156 | 0.157 | 0.148 (0.140) | 0.309 (0.130) |
| | | | 67% | 0.156 | 0.157 | 0.148 (0.186) | 0.475 (0.161) |
| | | | 76% | 0.156 | 0.157 | 0.161 (0.265) | 0.654 (0.188) |
| 3 | 0 | 0.05 | 0% | 0.721 | 0.750 | 0.716 (0.097) | 0.716 (0.097) |
| | | | 15% | 0.721 | 0.750 | 0.720 (0.106) | 0.844 (0.107) |
| | | | 30% | 0.721 | 0.750 | 0.725 (0.117) | 1.025 (0.119) |
| | | | 45% | 0.721 | 0.750 | 0.716 (0.139) | 1.294 (0.133) |
| | | | 60% | 0.721 | 0.750 | 0.716 (0.181) | 1.789 (0.168) |
| | | | 67% | 0.721 | 0.750 | 0.739 (0.255) | 2.247 (0.195) |

the above equations let us consider a third estimating equation which we write as

$$U_3(\beta) = \int \{\mathcal{Z}(t) - \mathcal{E}_\beta(Z|t)\} d\hat{F}(t) = 0. \tag{7.19}$$

Note that, upon defining the stochastic process $W(t) = \hat{S}(t)\{\sum_{i=1}^n Y_i(t)\}^{-1}$ we can rewrite (7.19) in the usual counting process terminology as

$$U_3(\beta) = \int W(t)\{\mathcal{Z}(t) - \mathcal{E}_\beta(Z|t)\} d\bar{N}(t) = 0.$$

For practical calculation note that $W(X_i) = \hat{F}(X_i+) - \hat{F}(X_i)$ at each observed failure time $X_i$, i.e., the jump in the KM curve. When there is no censoring, then clearly

$$U_1(\beta) = U_2(\beta) = U_3(\beta).$$

More generally $U_1(\beta)$ may not be available and solutions to $U_2(\beta) = 0$ and $U_3(\beta) = 0$ do not coincide or converge to the same population counterparts even under independent censoring. They would only ever converge to the same quantities under the unrealistic assumption that the data are exactly generated by a proportional hazards model. As argued in the previous section we can assume that this never really holds in practical situations.

Many other possibilities could be used instead of $U_3(\beta)$, ones in which other consistent estimates of $F(t)$ are used in place of $\hat{F}(t)$, for example, the Nelson-Aalen estimator or, indeed, any parametric estimate for marginal survival. If we were to take the route of parametric estimates of marginal survival, we would need to be a little cautious since these estimates could also contain information on the parameter $\beta$ which is our central focus. However, we could invoke a conditional argument, i.e., take the marginal survival estimate as fixed and known at its observed value or argue that the information contained is so weak that it can be ignored. Although we have not studied any of these we would anticipate the desirable properties described below to still hold. Stronger modelling assumptions are also possible (Moeschberger and Klein 1985, Klein et al. 1990).

Note also that the left-hand side of the equation is a special case of the weighted scores under the proportional hazards model (Harrington and Fleming 1982, Lin 1991, Newton and Raftery 1994). However those weighted scores were not proposed with the non-proportional hazards

model in mind, and the particular choice of $W(\cdot)$ used here was not considered in those papers. Indeed other choices for the weights will lead to estimators closer to the partial likelihood itself, in the sense that under a non-proportional hazards model and in the presence of censoring, the broader class of weighted estimates will not converge to quantities that remain unaffected by an independent censoring mechanism. On the other hand, the estimating equation based on $U_3$ is in the same spirit as the approximate likelihood of Oakes (1986) for censored data and the M-estimate of Zhou (1992) for censored linear models. Hjort (1992) also mentioned the use of the reciprocal of the Kaplan-Meier estimate of the censoring distribution as weights in parametric survival models, and these weights are the same as $W(\cdot)$ defined here. For the random effects model - a special case of this is the stratified model which, in turn, can be expressed in the form (6.2) - we can see, even when we know that (6.3) is severely misspecified, that we can still obtain estimates of meaningful quantities. The average effect resulting from the estimating equation $U_3$ is clearly of interest.

For the stratified model, $\mathcal{Z}(X_i)$ is contrasted with its expectation $\mathcal{E}_\beta(Z|X_i, s)$. Here, the inclusion of $s$ is used to indicate that if $Z_i$ belongs to stratum $s$ then the reference risk set for $\mathcal{E}_\beta(Z|X_i, s)$ is restricted to members of this same stratum. Note that for time-dependent $s(t)$ the risk set is dynamic, subjects entering and leaving the set as they become at risk. The usual estimating equation for stratified models is again of the form $U(\beta)$ and, for the same reasons as recalled above and described more fully in Xu and O'Quigley (1998) we might prefer to use

$$U_s(\beta) = \int \{\mathcal{Z}(t) - \mathcal{E}_\beta(Z|t, s)\} d\hat{F}(t) = 0\,. \tag{7.20}$$

Even weaker assumptions (not taking the marginal $F(t)$ to be common across strata) can be made and, at present, this is a topic that remains to be studied.

*Zeros of estimating equations*

Referring back to Section 7.4 we can immediately deduce that the zeros of the estimating equations provide consistent estimates of $\beta$ under the model. Below we consider zeros of the estimating equations when the model is incorrectly specified. This is important since, in practice, we can assume this to be the case. Most theoretical developments proceed

under the assumptions that the model is correct. We would have that $\hat{\beta}$ where $U_2(\hat{\beta}) = 0$ is consistent for $\beta$. Also $\tilde{\beta}$ where $U_3(\tilde{\beta}) = 0$ is consistent for a parameter of interest, namely the average effect. From the mean value theorem we write

$$U_2(\hat{\beta}) = U_2(\beta_0) + (\hat{\beta} - \beta_0) \left\{ \frac{\partial U_2(\beta)}{\partial \beta} \right\}_{\beta=\xi},$$

where $\xi$ lies strictly on the interior of the interval with endpoints $\beta_0$ and $\hat{\beta}$. Now $U_2(\hat{\beta}) = 0$ and $U_2'(\xi) = \sum_{i=1}^n \delta_i \mathcal{V}_\xi(Z|X_i)$ so that $\mathrm{Var}(\hat{\beta}) \approx 1/\sum_{i=1}^n \delta_i \mathrm{Var}(Z|X_i)$. This is the Cramer-Rao bound and so the estimate is a good one. Although the sums are of variables that we can take to be independent they are not identically distributed. Showing large sample normality requires verification of the Lindeburgh condition but, if awkward, this is not difficult. All the necessary ingredients are then available for inference. However, as our recommended approach, we adopt a different viewpoint based on the functional central limit theorem rather than a central limit theorem for independent variables. This is outlined in some detail in the following chapter.

*Large sample properties of solutions to estimating equations*

The reason for considering estimating equations other than (7.18) is because of large sample properties. Without loss of generality, for any multivariate categorical situation, a non-proportional hazards model (Equation 6.2) can be taken to generate the observations. Suppose that for this more general situation we fit the best available model, in particular the proportional hazards model (Equation 6.3). In fact, this is what always takes place when fitting the Cox model to data. It will be helpful to have the following definition:

**Definition 7.11** *The average conditional variance $A(\beta)$ is defined as;*

$$A(\beta) = \int_0^\infty \left\{ E_\beta(Z^2|t) - E_\beta^2(Z|t) \right\} dF(t).$$

Note that the averaging does not produce the marginal variance for that we would need to include a further term which measures the variance of the conditional expectations. Under the conditions on the censoring of Breslow and Crowley (1974), essentially requiring that, for each $t$, as $n$ increases, the information increases at the same rate, then $nW(t)$ converges in probability to $w(t)$. Under these same conditions,

recall that the probability limit as $n \to \infty$ of $\mathcal{E}_\beta(Z|t)$ under model (6.2) is $E_\beta(Z|t)$, that of $\mathcal{E}_\beta(Z^2|t)$ is $E_\beta(Z^2|t)$ and that of $\mathcal{V}_\beta(Z|t)$ is $V_\beta(Z|t)$. The population conditional expectation and variance, whether the model is correct or not, are denoted by $E(Z|t)$ and $V(Z|t)$, respectively. We have an important result due to Struthers and Kalbfleisch (1986).

**Theorem 7.2** *Under model 6.2 the estimator $\hat{\beta}$, such that $U_2(\hat{\beta}) = 0$, converges in probability to the constant $\beta_{PL}$, where $\beta_{PL}$ is the unique solution to the equation*

$$\int_0^\infty w^{-1}(t)\left\{E(Z|t) - E_\beta(Z|t)\right\} dF(t) = 0, \tag{7.21}$$

*provided that $A(\beta_{PL})$ is strictly greater than zero.*

Should the data be generated by model (6.3) then $\beta_{PL} = \beta$, but otherwise the value of $\beta_{PL}$ would depend upon the censoring mechanism in view of its dependence on $w(t)$. Simulation results below on the estimation of average effect show a very strong dependence of $\beta_{PL}$ on an independent censoring mechanism. Of course, under the unrealistic assumption that the data are exactly generated by the model, then, for every value of $t$, the above integrand is identically zero, thereby eliminating any effect of $w(t)$. In such situations the partial likelihood estimator is more efficient and we must anticipate losing efficiency should we use the estimating equation $U_3(\beta)$ rather than the estimating equation $U_2(\beta)$.

Viewing the censoring mechanism as a nuisance feature of the data we might ask the following question: were it possible to remove the censoring then to which population value do we converge? We would like an estimating equation that, in the presence of an independent censoring mechanism, produces an estimate that converges to the same quantity we would have converged to had there been no censoring. The above estimating equation (7.19) has this property. This is summarized in the following theorem of Xu and O'Quigley (1998), which is an application of Theorem 3.2 in Lin (1991).

**Theorem 7.3** *Under model 6.2 the estimator $\tilde{\beta}$, such that $U_3(\tilde{\beta}) = 0$, converges in probability to the constant $\beta^*$, where $\beta^*$ is the unique solution to the equation*

$$\int_0^\infty \left\{E(Z|t) - E_\beta(Z|t)\right\} dF(t) = 0, \tag{7.22}$$

*provided that $A(\beta^*)$ is strictly greater than zero.*

None of the ingredients in the above equation depends on the censoring mechanism. In consequence the solution itself, $\beta = \beta^*$, is not influenced by the censoring. Thus the value we estimate in the absence of censoring, $\beta^*$, is the same as the value we estimate when there is censoring. A visual inspection of equations (7.21) and (7.22) suffices to reveal why we argue in favor of (7.19) as a more suitable estimating equation than (7.18) in the presence of non proportional hazard effects. Furthermore, the solution to (7.19) can be given a strong interpretation in terms of average effects. We return to this in more detail, but we can already state a compelling argument for the broader interpretability of $\beta^*$.

## 7.6 Consistency and asymptotic normality of $\tilde{\boldsymbol{\beta}}$

We have that $\mathcal{E}_\beta(Z|t) = S^{(1)}(\beta,t)/S^{(0)}(\beta,t)$, and that $W(t) = \hat{S}(t)/\{nS^{(0)}(0,t)\}$. Under an independent censoring mechanism, $s^{(1)}(\beta(t),t)/s^{(0)}(\beta(t),t) = E\{Z(t)|T=t\}$, and $s^{(1)}(\beta,t)/s^{(0)}(\beta,t)$ is what we get when we impose a constant $\beta$ through time in place of $\beta(t)$, both of which do not involve the censoring distribution. In addition $v(t) = v(\beta(t),t) = \mathrm{Var}\{Z(t)|T=t\}$. We take it that $nW(t)$ converges in probability to a non-negative bounded function $w(t)$ uniformly in $t$. Then we have $w(t) = S(t)/s^{(0)}(0,t)$. Using the same essential approach as that of Andersen and Gill (1982) it is seen, under the model and an independent censoring mechanism, that the marginal distribution function of $T$ can be written

$$F(t) = \int_0^t w(t)s^{(0)}(\beta(t),t)\lambda_0(t)dt. \tag{7.23}$$

**Theorem 7.4** *Under the non-proportional hazards model and an independent censorship the estimator $\tilde{\beta}$ converges in probability to the constant $\beta^*$, where $\beta^*$ is the unique solution to the equation*

$$\int_0^\infty \left\{ \frac{s^{(1)}(\beta(t),t)}{s^{(0)}(\beta(t),t)} - \frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)} \right\} dF(t) = 0, \tag{7.24}$$

*provided that $\int_0^\infty v(\beta^*,t)dF(t) > 0$.*

It is clear that equation (7.24) does not involve censoring. Neither then does the solution to the equation, $\beta^*$. As a contrast the maximum partial likelihood estimator $\hat{\beta}_{PL}$ from the estimating equation $U_2 = 0$ converges to the solution of the equation

$$\int_0^\infty \left\{ \frac{s^{(1)}(\beta(t), t)}{s^{(0)}(\beta(t), t)} - \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} \right\} s^{(0)}(\beta(t), t)\lambda_0(t)dt = 0. \qquad (7.25)$$

This result was obtained by Struthers and Kalbfleisch (1986). Should the data be generated by the proportional hazards model, then the solutions of (7.24) and (7.25) are both equal to the true regression parameter $\beta$. In general, however, these solutions will be different, the solution to (7.25) depending on the unknown censoring mechanism through the factor $s^{(0)}(\beta(t), t)$. The simulation results of Table 7.1 serve to underline this fact in a striking way. The estimate $\tilde{\beta}$ can be shown to be asymptotically normal with mean zero and variance that can be written down. The expression for the variance is nonetheless complicated and is not reproduced here since it is not used. Instead we base inference on functions of Brownian motion as described in the next chapter.

## 7.7 Interpretation for $\beta^*$ as average effect

The solution $\beta^*$ to the large sample equivalent to the estimating equation $U_3(\beta)$, i.e., Equation 7.24 can be viewed as an average regression effect. In the equation $s^{(1)}(\beta(t), t)/s^{(0)}(\beta(t), t) = E\{Z(t)|T = t\}$, and $s^{(1)}(\beta^*, t)/s^{(0)}(\beta^*, t)$ results when $\beta(t)$ is restricted to be a constant; the difference between these two is zero when integrated out with respect to the marginal distribution of failure time. Suppose, for instance, that $\beta(t)$ decreases over time, then earlier on $\beta(t) > \beta^*$ and $s^{(1)}(\beta(t), t)/s^{(0)}(\beta(t), t) > s^{(1)}(\beta^*, t)/s^{(0)}(\beta^*, t)$; whereas later we would have the opposite effect whereby $\beta(t) < \beta^*$ and $s^{(1)}(\beta(t), t)/s^{(0)}(\beta(t), t) < s^{(1)}(\beta^*, t)/s^{(0)}(\beta^*, t)$. We can write, $v(\beta, t) = \partial/\partial\beta\{s^{(1)}(\beta, t)/s^{(0)}(\beta, t)\}$ and, applying a first-order Taylor series approximation to the integrand of (7.24), we have

$$\int_0^\infty v(t)\{\beta(t) - \beta^*\}dF(t) \approx 0, \qquad (7.26)$$

where $v(t) = v(\beta(t), t) = \text{Var}\{Z(t)|T = t\}$. Therefore

$$\beta^* \approx \frac{\int_0^\infty v(t)\beta(t)dF(t)}{\int_0^\infty v(t)dF(t)} \tag{7.27}$$

is a weighted average of $\beta(t)$ over time. According to Equation 7.27 more weights are given to those $\beta(t)$'s where the marginal distribution of $T$ is concentrated, which simply means that, on average, we anticipate there being more individuals subjected to those particular levels of $\beta(t)$. The approximation of Equation 7.27 also has an interesting connection with Murphy and Sen (1991), where they show that if we divide the time domain into disjoint intervals and estimate a constant $\beta$ on each interval, in the limit as $n \to \infty$ and the intervals become finer at a certain rate, the resulting $\hat{\beta}(t)$ estimates $\beta(t)$ consistently. In their large sample studies, they used a (deterministic) piecewise constant parameter $\bar{\beta}(t)$, which is equivalent to Equation 7.27 restricted to individual intervals. They showed that $\bar{\beta}(t)$ is the best approximation to $\hat{\beta}(t)$, in the sense that the integrated squared difference $\int\{\hat{\beta}(t) - \bar{\beta}(t)\}^2 dt \to 0$ in probability as $n \to \infty$, at a faster rate than any other choice of such piecewise constant parameters. In Equation (7.27) if $v(t)$, the conditional variance of $Z(t)$, changes relatively little with time apart from for large $t$, when the size of the risk sets becomes very small, we can make the approximation $v(t) \equiv c$ and it follows that

$$\beta^* \approx \int_0^\infty \beta(t)dF(t) = E\{\beta(T)\}. \tag{7.28}$$

In practice, $v(t)$ will often be approximately constant, an observation supported by our own practical experience as well as with simulated data sets. For a comparison of two groups coded as 0 and 1, the conditional variance is of the form $p(1-p)$ for some $0 < p < 1$, and this changes relatively little provided that, throughout the study, $p$ and $1-p$ are not too close to zero. The approximate constancy of this conditional variance is used in the sample size calculation for two-group comparisons (Kim and Tsiatis 1990). In fact, we only require the weaker condition that $\mathrm{Cov}(v(T), \beta(T)) = 0$ to obtain Equation 7.28, a constant $v(t)$ being a special case of this. Even when this weaker condition does not hold exactly, $\int \beta(t)dF(t)$ will still be close to $\beta^*$.

Xu and O'Quigley (1998) carried out simulations to study the approximation of $\int \beta(t)dF(t)$ to $\beta^*$. Some of those findings are shown in Table 7.1 and these are typical of the findings from a wide variety of other situations. The results are indeed striking. It is also most likely

true that it is not well known just how strong is the dependence of the partial likelihood estimator on an independent censoring mechanism when the data are generated by a non-proportional hazards model. Since, in practical data analysis, such a situation will almost always hold, we ought be rather more circumspect about the usual estimators furnished by standard software.

In the table the data are simulated from a simple two-step time-varying regression coefficients model, with baseline hazard $\lambda_0(t) = 1$, $\beta(t) = \beta_1$ when $t < t_0$ and $\beta_2$ otherwise. The covariate $Z$ is distributed as Uniform(0,1). At time $t_0$ a certain percentage of subjects at risk are censored. The value $\hat{\beta}_{PL}$ is the partial likelihood estimate when we fit a proportional hazards model to the data. Table 7.1 summarizes the results of 200 simulations with sample size of 1600. We see that $\int \beta(t)dF(t)$ is always close to $\beta^*$, for the values of $\beta$ that we might see in practice. The most important observation to be made from the table is the strong dependence of $\hat{\beta}_{PL}$ on an independent censoring mechanism, the value to which it converges changing substantially as censoring increases. The censoring mechanism here was chosen to emphasize the difference between $\hat{\beta}_{PL}$ and $\tilde{\beta}$, since $\tilde{\beta}$ puts (asymptotically) the correct weights on the observations before and after $t_0$. In other cases the effect of censoring may be weaker. Nonetheless, it is important to be aware of the behavior of the partial likelihood estimator under independent censoring and non-proportional hazards and the subsequent difficulties in interpreting the partial likelihood estimate in general situations.

The bracketed figures in Table 7.1 give the standard errors of the estimates from the simulations. From these we can conclude that any gains in efficiency of the partial likelihood estimate can be very quickly lost to biases due to censoring. When there is no censoring the estimators are the same. As censoring increases we see differences in the standard errors of the estimates, the partial likelihood estimate being more efficient; but we also see differences in the biases. Typically, these latter differences are at least an order of magnitude greater.

## 7.8 Exercises and class projects

1. Show that, under an independent censoring mechanism, $\hat{H}(t)$, as defined in Section 7.3, provides a consistent estimate of $H(t)$.

2. Show that the variance expression, $V(\beta, t)$, using the Andersen and Gill notation (see Section 7.3) is the same as $\mathcal{V}_\beta(Z|t)$ using the notation of Section 7.4. Explain why $\mathrm{Var}(Z|t)$ is consistently estimated by $\mathcal{V}_{\hat{\beta}}(Z|t)$ but that $\mathrm{Var}(Z|t)$ is not generally equal to $v(\beta, t)$.

3. For the general model, suppose that $\beta(t)$ is linear so that $\beta(t) = \alpha_0 + \beta t$. Show that $\mathcal{E}_{\beta(t)}(Z^k|t)$ does not depend upon $\alpha_0$.

4. Sketch an outline of a proof that $\mathrm{Var}(Z|t)$ is consistently estimated by $\mathcal{V}_{\hat{\beta}}(Z|t)$ and that $E\,\mathrm{Var}(Z|t)$ is consistently estimated by $E\,\mathcal{V}_{\hat{\beta}}(Z|t)$.

5. As for the previous question, indicate why $\int \mathcal{V}_{\hat{\beta}}(Z|t)d\hat{F}(t)$ would be consistent for $E\,\mathrm{Var}(Z|t)$.

6. Show that $\mathcal{V}_\beta(Z|t) = \partial\,\mathcal{E}_\beta(Z|t)/\partial\beta$ and identify the conditions for the relationship; $\int \mathcal{V}_\beta(Z|t) = \int \partial\,\mathcal{E}_\beta(Z|t)/\partial\beta = \partial\left\{\int\,\mathcal{E}_\beta(Z|t)\right\}/\partial\beta$ to hold.

7. Consider some parametric non proportional hazards model (see Chapter 4), in which the conditional density of $T$ given $Z = z$ is expressed as $f(t|z)$. Suppose the marginal distribution of $Z$ is $G(z)$. Write down estimating equations for the unknown parameters based on the observations $Z_i$ at the failure times $X_i$.

8. Use some data set to fit the proportional hazards model. Estimate the parameter $\beta$ on the basis of estimating equations for the observations $Z_i^2$ rather than $Z_i$. Derive another estimate based on estimating equations for $\sqrt{Z}_i$. Compare the estimates.

9. Write down a set of estimating equations based on the observations, $Z_i^p$ , $p > 0\,, i = 1, \ldots, n$. Index the estimate $\hat{\beta}$ by $p$, i.e., $\hat{\beta}(p)$. For a given data set, plot $\hat{\beta}(p)$ as a function of $p$.

10. Use analytical or heuristic arguments to described the expected behavior of $\hat{\beta}(p)$ as a function of $p$ under (1) data generated under a proportional hazards model, (2) data generated under a non-proportional hazards model where the effect declines monotonically with time.

11. Consider a proportional hazards model in which we also know that the marginal survival is governed by a distribution $F(t;\theta)$ where $\theta$ is

not known. Suppose that it is relatively straightforward to estimate $\theta$, by maximum likelihood or by some graphical technique. Following this we base an estimating equation for the unknown regression coefficient, $\beta$, on $U(\beta|\hat{\theta}) = \int \{\mathcal{Z}(t) - \mathcal{E}_\beta(Z|t)\} dF(t;\hat{\theta})$. Comment on this approach and on the properties you anticipate it conferring on the estimate $\hat{\beta}$.

12. Use the approach of the preceding question on some data set by (1) approximating the marginal distribution by an exponential distribution, (2) approximating the marginal distribution by a log-normal distribution.

13. Using again the approach of the previous two questions show that, if the proportional hazards models is correctly specified then the estimate $\hat{\beta}$ based on $F(t;\theta)$ is consistent whether or not the marginal model $F(t;\theta)$ is correctly specified.

14. Supposing that the function $\beta(t)$ is linear so that $\beta(t) = \alpha_0 + \beta t$. Show how to estimate the function $\beta(t)$ in this simple case. Note that we can use this model to base a test of the proportional hazards assumption via a hypothesis test that $H_0 : \beta = 0, \alpha_0 \neq 0$ (Cox 1972).

15. Investigate the assertion that it is not anticipated for $v(t)$, the conditional variance of $Z(t)$, to change much with time. Use the model-based estimates of $v(t)$ and different data sets to study this question informally.

16. In epidemiological studies of breast cancer it has been observed that the tumor grade is not well modeled on the basis of a proportional hazards assumption. A model allowing a monotonic decline in the regression coefficient $\beta(t)$ provides a better fit to observed data. On the basis of observations some epidemiologists have argued that the disease is more aggressive (higher grade) in younger women. Can you think of other explanations for this observed phenomenon?