New Approaches in Applied Statistics
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 16, Ljubljana: FDV, 2000

Inference for the Cox Model under Proportional and Non-Proportional Hazards

John O'Quigley[1] and Ronghui Xu[2]

Abstract

We consider some new ideas concerning inference in the proportional hazards model and the broader non proportional hazards model. Recall that bivariate regression models generally focus in an explicit way on the conditional distribution of one variable given the other. There are always two ways of doing this but, typically, it is most natural to condition on the explanatory or design variable. However, as far as concerns inference for proportional hazards regression, it is in fact more natural to condition the other way around. This simple observation leads to many results. Among these are a natural estimator of average effects under non-proportional hazards, a straightforward and natural way to assess fit and a new estimator of the survivorship function conditonal on some set of covariates. These ideas all stem from a main theorem which we describe below. All the ideas for the bivariate case, i.e. a single explanatory variable in the model, generalize readily to the multivariate case.

1 Introduction

Few statistical techniques have had as great an impact on the applied medical and biological sciences as the proportional hazards model (Cox, 1972). It has become difficult to find a medical journal in which the method is not referred to at least once, and usually several, times during the year. In a review article (Andersen, 1991) it was stated that the annual rate of citation of Cox's paper had risen to around 540 by the end of the 1980s, the majority of which were in medical journals,

[1] Department of Mathematics, University of California at San Diego, La Jolla, CA 92093, USA.
[2] Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA.

a statement leading the author to conclude that "Cox's paper has had an enormous impact on medical research". Since Andersen's review the popularity of the model has continued to increase, finding new applications and extensions in fields as varied as economics, sociology and engineering.

The partial likelihood estimator for the regression coefficient β in the model was first shown to be consistent by Cox (1975) and later, more formally, by Tsiatis (1981). Consistency was established using a martingale approach by Andersen and Gill (1982). All these papers indicate that consistency is maintained under an independent censoring mechanism conditional on the covariate. In more general contexts than survival analysis modelling, consistency often holds under broader assumptions than those of the working model, the population value to which the estimator converges coinciding with that for the restricted model when such a model is correct. The most immediate example of this would be the maximum likelihood estimator of the mean for i.i.d. normal variates, converging to the mean for distributions other than normal, providing the mean exists.

For circumstances in which the regression coefficient exhibits some time dependency, expressed as $\beta(t)$, the partial likelihood estimator converges to some population parameter depending in a complex way on the underlying censoring mechanism. This property limits the practical application of the estimator. So, although the partial likelihood estimator is consistent under the proportional hazards model, it is not consistent for any meaningful parameter, i.e. one that does not involve censoring, under broader models in which the regression effect $\beta(t)$ is not constant through time. Indeed, the effect of censoring on the partial likelihood estimate can be considerable (Xu, 1996), whereas, for the estimator proposed here, not only does censoring not impact the population parameter to which we converge but also the parameter can be given a concrete interpretation as "average effect". The proposed estimator can be seen to be consistent in the more usual situation in which the data are generated by a mechanism that is not exactly equal to but only approximated by the working model in which hazard ratios are taken to be constant.

For any regression model, attention focuses on the conditional distribution of a response variable given the explanatory variables. For survival models we usually view the response variable as time and the explanatory variables as associated prognostic measurements. In the Cox model, reliance upon inferential procedures which are invariant to monotonic time transformation, means that attention focuses more naturally upon the conditional distribution of the explanatory variables given time rather than the other way around. This observation is implicit in much of the work carried out on the model and we make it explicit here via Theorem 1. It is our view that this theorem is very important in understanding the model. Sections 2.1 and

2.2 outline this reasoning.

This theorem is central to understanding inference in the Cox model. On the basis of it we can see how to derive the estimate of average effect. But also a very straightforward way to assess whether or not an assumption of proportional hazards is reasonable. Another application of theorem 1 concerns the conditional distribution of survival given that the explanatory variables belong to some set. Solutions have already been proposed for this question. However it follows from the theorem that an alternative, in many ways more natural, solution is easily available. The same theorem can also be used to examine the overall fit of a proportional hazards model.

2 Models and conditional distributions

The applied problem concerns the detection and quantification of association between a randomly chosen subject's survival time and some explanatory variable, possibly in the presence of random censoring. So, in a survival study with n subjects, let $T_1, T_2, ..., T_n$ be the failure times, and $C_1, C_2, ..., C_n$ be the censoring times for the individuals $i = 1, 2, ..., n$. For each i we observe $X_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$, where $I(\cdot)$ is the indicator function. Define the "at risk" indicator $Y_i(t) = I(X_i \geq t)$. We will also use the counting process notation: let $N_i(t) = I\{T_i \leq t, T_i \leq C_i\}$ and $\bar{N}(t) = \sum_1^n N_i(t)$. The inverse of the function $\bar{N}$ written $\bar{N}^{-1}(j)\,(j = 1, 2, ...)$, is defined to be $\inf t : \bar{N}(t) = j$. The total number of observed failures is k so that $k = \bar{N}(\infty)$. The Kaplan-Meier estimate of survival is denoted $\hat{S}(t)$ and the Kaplan-Meier estimate of the distribution function by $\hat{F}(t) = 1 - \hat{S}(t)$. Usually we are interested in the situation where each subject has related covariates, or explanatory variables, Z_i $(i = 1, 2, ..., n)$. All of the results given here hold for an independent censorship model, a common assumption in survival studies. Many of the results still hold or can easily be extended to apply under the weaker condition of a conditional independent censorship model. Mostly, for ease of exposition, we assume the covariate Z to be one dimensional. Z in general could be time-dependent, in which case it is assumed to be a predictable stochastic process and we will use the notation $Z(t)$, $Z_i(t)$, etc.

2.1 Conditional distribution of T given Z

The Cox (1972) proportional hazards model assumes that the hazard function $\lambda_i(t)$ $(i = 1, ..., n)$ for individuals with different covariates, $Z_i(t)$, can be written

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta Z_i(t)\}, \tag{2.1}$$

where $\lambda_0(t)$ is a fixed unknown "baseline" hazard function, and β is a relative risk parameter to be estimated. Statistical inference on β is traditionally carried out through maximizing Cox's (1975) partial likelihood

$$L(\beta) = \prod_{i=1}^{n} \pi_i(\beta, X_i)^{\delta_i}, \tag{2.2}$$

where

$$\pi_i(\beta, t) = K(t) Y_i(t) \exp\{\beta Z_i(t)\} \tag{2.3}$$

and $K(t)$ standardizes the $\pi_i(\beta, t)$ to be probabilities, i.e.

$$K^{-1}(t) = \sum_{\ell=1}^{n} Y_\ell(t) \exp\{\beta Z_\ell(t)\}$$

Indeed, under (2.1), $\pi_i(\beta, t)$ is exactly the conditional probability that at time t, it is precisely individual i who is selected to fail, given all the individuals at risk and given that one failure occurs. Let

$$\mathcal{E}_\beta(Z|t) = \sum_{\ell=1}^{n} Z_\ell(t)\pi_\ell(\beta, t) = \sum_{\ell=1}^{n} K(t) Y_\ell(t) Z_\ell(t) \exp\{\beta Z_\ell(t)\}. \tag{2.4}$$

As noted by Andersen and Gill (1982), $\mathcal{E}_\beta(Z|t)$ is the expectation of the covariate $Z(t)$ with respect to the probability distribution $\{\pi_i(\beta, t)\}_i$. Taking the logarithm and derivative in (2.2) with respect to β, we obtain the score function

$$U(\beta) = \sum_{i=1}^{n} \delta_i r_i(\beta), \quad \text{where} \quad r_i(\beta) = Z_i(X_i) - \mathcal{E}_\beta(Z|X_i)\,, \ (i = 1, ..., n). \tag{2.5}$$

Setting (2.5) equal to zero, we get the maximum partial likelihood estimate (MPLE) $\hat{\beta}$. Note that the $r_i(\hat{\beta})$ are the Schoenfeld (1982) residuals at each observed failure time X_i. Also the expectation $\mathcal{E}_\beta(Z|X_i)$ is worked out with respect to an exponentially tilted distribution, the stronger the regression effects the greater the tilting and the more the mean is shifted away from the empirical mean.

In practice the proportional hazards assumption may fail to be met in a number of ways (Schoenfeld, 1980; Lagakos and Schoenfeld, 1984). Inadequacies in the chosen covariate structure or time dependent effects will result in non-proportional hazards. What is more, unlike the set up for normal linear regression, we even have a theoretical impossibility of exactly meeting the assumption of proportional hazards for nested models. This latter example has been reported in a number of studies since the early 80's (Lancaster and Nickell, 1980; Gail et al., 1984; Struthers and Kalbfleisch, 1986; Bretagnolle and Huber-Carol, 1988; Anderson and Fleming, 1995; Ford et al., 1995). In fact, suppose that we have

$$\lambda(t|Z_1, Z_2) = \lambda_0(t) \exp\{\beta_1 Z_1 + \beta_2 Z_2\}. \tag{2.6}$$

where Z_1, Z_2 are vectors of time-invariant covariates, then

$$S(t|Z_1, Z_2) = S_0(t)^{\exp\{\beta_1 Z_1 + \beta_2 Z_2\}},$$

where $S_0(t)$ is the survivorship function corresponding to $\lambda_0(t)$. The survivorship function $S(t|Z_1, Z_2)$ is of the right form to belong to the proportional hazards class. However, for some partition H over the domain of definition of Z_2 and where $p(\cdot)$ associates probabilities to the subsets $z_2 \in H$, it follows by the law of total probability that;

$$S(t|Z_1) = \sum_{z_2 \in H} S(t|Z_1, z_2)p(z_2)$$

so that, generally, we would not anticipate $S(t|Z_1)$ to also belong to the proportional hazards class. In consequence were we to fit a submodel to data generated via a broader model then we anticipate inconsistency (sometimes referred to as asymptotic bias) in the estimates. This has attracted particular attention in the context of Cox regression where authors have pointed out the dangers of overlooking covariables which may be associated with survival. We do not see this viewpoint as being very helpful and prefer simply to view all models as approximations to more complex realities.

We can nonetheless find a simple expression for this more complex reality. For a single binary covariate the true mechanism generating the data can be written, without loss of generality, as

$$\lambda_i(t|Z(t)) = \lambda_0(t) \exp\{\beta(t)Z_i(t)\}, \tag{2.7}$$

We can describe this model as a non proportional hazards model in order to situate it alongside the proportional hazards model in which $\beta(t) = \beta$, a constant. However it is more a representation of reality than a model since no restriction at all, apart from the necessary one of being positive, is imposed on $\lambda_i(t|Z(t))$. The same level of generality also holds for more complicated situations than that of the simple binary covariate, for example p groups being represented by $p-1$ indicator variables.

For continuous covariates full generality no longer holds but the above model would still be richer than (2.1). Also by dividing continuous covariates into categories the model can be made as general as we wish, the only limitation being sample size and sparse cells. It is then helpful to see the above expression as a representation of reality, the proportional hazards model then imposing a well defined restriction upon this reality. The relevant question, when fitting a model in which β is not allowed to vary through time, is what meaning can we give to such a β.

Models making weaker restrictions on $\beta(t)$ than being constant for all t have been looked at by Moreau et al. (1985), O'Quigley and Pessione (1989, 1991),

Liang et al. (1990), Zucker and Karr (1990), Murphy and Sen (1991), Gray (1992), Hastie and Tibshirani (1993), Verweij and Van Houwelingen (1995), Lausen and Schumacher (1996), Marzec and Marzec (1997), and references therein. The main emphasis of these papers was to estimate the regression effect $\beta(t)$ as a function of t, under the particular model chosen. Other approaches, in particular smoothing techniques, can also be used to estimate $\beta(t)$. Such analyses can be involved and we limit ourselves here to the simpler question of estimation of average effects without necessarily investigating the whole of $\beta(t)$ through time.

2.2 Conditional distribution of Z given T

Although most studies, whether following some experimental design or not, view Z as fixed and T as being random, it turns out to be very helpful to also consider fixing certain values of T, notably the observed failure times, and to study the distribution of Z at these times. The following theorem provides a central result.

Theorem 1
Under model (2.8), and where $\hat{\beta}(t)$ is any consistent estimate of $\beta(t)$, the conditional distribution function of $Z(t)$ given $T = t$ and $C > t$ is consistently estimated by

$$\hat{F}_t(z|t) \;=\; \hat{P}(Z(t) \leq z | T = t, C > t) \;=\; \sum_{\{\ell : Z_\ell(t) \leq z\}} \pi_\ell\{\hat{\beta}(t), t\}.$$

Proof: See Xu (1996).

In the light of the above theorem, and letting

$$\mathcal{E}_\beta(Z^k|t) = \sum_{i=1}^{n} Z_i^k(t)\pi_i\{\beta(t), t\} = \sum_{\ell=1}^{n} K(t)Y_\ell(t)Z_\ell^k(t)\exp\{\beta(t)Z_\ell(t)\}\,,\;\; \ell = 1, 2, .. \quad (2.8)$$

we have that $\mathcal{E}_{\hat{\beta}}(Z^k|t)$ provide consistent estimates of $E(Z^k(t)|T = t, C > t)$, under the assumptions. In particular a consistent estimate of the variance $\mathrm{Var}(Z(t)|T = t, C > t)$, is $\mathcal{V}_{\hat{\beta}}(Z|t)$ where

$$\mathcal{V}_\beta(Z|t) = \mathcal{E}_\beta(Z^2|t) - \mathcal{E}_\beta^2(Z|t) \quad (2.9)$$

The multivariate generalization is straightforward. To see this consider for example a model with two covariates,

$$\lambda_i(t) = \lambda_0(t)\exp\{\beta_1(t)Z_{1i}(t) + \beta_2(t)Z_{2i}(t)\}. \quad (2.10)$$

We would then have

$$\pi_i\{\beta_1(t), \beta_2(t), t\} = K(t)Y_i(t)\exp\{\beta_1(t)Z_{1i}(t) + \beta_2(t)Z_{2i}(t)\}\,, \quad (2.11)$$

where $K(t)$ is essentially the same as in (2.3) generalized in an obvious way. This then leads to expressions such as

$$\mathcal{E}_{\beta_1\beta_2}(Z_1^k|t) = \sum_{i=1}^{n} Z_{1i}^k(t)\pi_i\{\beta_1(t), \beta_2(t), t\}\,, \quad k = 1, 2, ..., \tag{2.12}$$

$$\mathcal{E}_{\beta_1\beta_2}(Z_1 Z_2|t) = \sum_{i=1}^{n} Z_{1i}(t)Z_{2i}(t)\pi_i\{\beta_1(t), \beta_2(t), t\}\,, \tag{2.13}$$

as consistent estimates, conditional upon $(T = t, C > t)$, for the marginal moments and cross product terms respectively.

3 Inference for the regression coefficient

3.1 Estimating equations

Firstly introduce the function $\mathcal{Z}(t)$, a step function with discontinuities at the points X_i, $i = 1, ..., n$, where it 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, an estimating equation arising as the derivative of the log partial likelihood, is;

$$U_1(\beta) = \int \{\mathcal{Z}(t) - \mathcal{E}_\beta(Z|t)\} dF_n(t) = 0 \tag{3.1}$$

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 $F_n(t) \xrightarrow{\text{p}} F(t)$. Since the increments, $1/n$, in the above equation are all the same size, we can cancel them and rewrite the equation as;

$$U_2(\beta) = \int \{\mathcal{Z}(t) - \mathcal{E}_\beta(Z|t)\} d\bar{N}(t) = 0 \tag{3.2}$$

which is now the more classic representation in this context, being expressed in terms of the counting processes $N_i(t)$. In the presence of censoring it is the above equation that is used to define the partial likelihood estimator, especially since $F_n(t)$ is no longer available and thereby $U_1(\beta)$ undefined. Before discussing the above equations let us consider a third estimating equation which we write as;

$$U_3(\beta) = \int \{\mathcal{Z}(t) - \mathcal{E}_\beta(Z|t)\} d\hat{F}(t) = 0 \tag{3.3}$$

Note that, upon defining the predictable stochastic process $W(t)$ where

$$W(t) = \frac{\hat{S}(t)}{\sum_{i=1}^{n} Y_i(t)}$$

we can rewrite (3.3) in the usual counting process terminology as

$$U_3(\beta) = \int W(t)\{\mathcal{Z}(t) - \mathcal{E}_\beta(Z|t)\}d\bar{N}(t) = 0. \tag{3.4}$$

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 (see below). 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 estimate. Although we have not studied any of these we would anticipate the desirable properties described in the next section to still hold.

3.2 Large sample properties

The reason for considering estimating equations other than (3.2) is because of large sample properties. Without loss of generality, for any multivariate categorical situation, model (2.7) can be taken to generate the observations. Suppose that for this more general situation we fit model (2.1). In fact this is what always takes place when fitting the Cox model to data. 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 denote the probability limit as $n \to \infty$ of $\mathcal{E}_\beta(Z|t)$ under model (2.7) by $E_\beta(Z|t)$, that of $\mathcal{E}_\beta(Z^2|t)$ by $E_\beta(Z^2|t)$ and that of $\mathcal{V}_\beta(Z|t)$ by $V_\beta(Z|t)$. The population conditional expectation and variance, whether the model is correct or not, are denoted by $E(Z|t)$ and $V(Z|t)$ respectively.

The maximum partial likelihood estimator $\hat{\beta}$ from the score function (3.2) was shown by Struthers and Kalbfleisch (1986) to converge to the population value β_{PL} which solves the equation

$$\int_0^\infty w^{-1}(t)\left\{E(Z|t) - E_\beta(Z|t)\right\} dF(t) = 0, \tag{3.5}$$

Should the data be generated by model (2.1) then $\beta_{PL} = \beta$, but otherwise the value of β_{PL} would depend upon the censoring mechanism in view of its dependance upon $w(t)$. Simulation results (Xu, 1996) show a strong dependence of β_{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 efficient and we must anticipate losing efficiency should we use different weights as in equation (3.3).

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 (3.3) 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 2

Under model (2.8) the estimator $\tilde{\beta}$, such that $U_3(\tilde{\beta}) = 0$, converges in probability to the constant β^*, where β^* is the unique solution to the equation

$$\int_0^\infty \{E(Z|t) - E_\beta(Z|t)\}\, dF(t) = 0, \tag{3.6}$$

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

$$A(\beta) = \int_0^\infty \left\{E_\beta(Z^2|t) - E_\beta^2(Z|t)\right\} dF(t), \tag{3.7}$$

None of the ingredients in the above two equations depend upon 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, β^*, is the same as the value we estimate when there is censoring. A visual inspection of equations (3.5) and (3.6) suffices to reveal why we argue in favor of (3.3) as a more suitable estimating equation than (3.2) in the presence of non proportional hazard effects. Furthermore the solution to (3.3) can be given a strong interpretation in terms of average effects. This is described in the following paragraph.

3.3 An interpretation for β^* as average effect

Since we can consider the full model 2.6 as generating the observations, then this implies that, for every t, there exists some value of $\beta(t)$ such that $E(Z|t) = E_{\beta(t)}(Z|t)$. On the basis of this we can write;

$$E(Z|t) = E_{\beta^*}(Z|t) + \{\beta(t) - \beta^*\}\left\{\frac{\partial E_\beta(Z|t)}{\partial \beta}\right\}_{\beta=\beta^*} + \{\beta(t) - \beta^*\}^2\left\{\frac{\partial^2 E_\beta(Z|t)}{\partial \beta^2}\right\}_{\beta=\xi} \tag{3.8}$$

where ξ lies strictly between $\beta(t)$ and β^*. Using the above and recalling the definition of β^* we have, as a Taylor series approximation

$$\int \{\beta(t) - \beta^*\} V_{\beta^*}(Z|t) dF(t) \approx 0. \tag{3.9}$$

Rearranging the above formula we have that

$$\beta^* \approx \frac{\int \beta(t) V_{\beta^*}(Z|t) dF(t)}{\int V_{\beta^*}(Z|t) dF(t)}. \tag{3.10}$$

and, if the conditional variance of Z changes relatively little with time, or depends on time such that $\text{Cov}\{\beta(T), V(Z|T)\} = 0$ then we can make the further approximation;

$$\beta^* \approx \int \beta(t) dF(t). \tag{3.11}$$

Our experience with some data sets as well as a large number of simulations (Xu 1996) suggest that this approximation ought work well in many practical situations. This enables a concrete interpretation to be given to β^* as average effect over some time interval upon which $\beta(t)$ is defined. Since it is also helpful to estimate to what extent $\beta(t)$ is changing over the interval, we can reason in an analagous way to the above to obtain:

$$\int \{\beta(t) - \beta^*\}^2 dF(t) \approx \int \left\{ \frac{E(Z - E_{\beta^*}(Z|t)|t)}{V_{\beta^*}(Z|t)} \right\}^2 dF(t) \tag{3.12}$$

The size of $\beta(t)$ and thereby β^* depend on the units of Z and it would be helpful to have some measure of the extent of departure from proportionality that does not depend on the units. The ϕ coefficient defined via:

$$\phi = \log \int V_{\beta^*}(Z|t) \{\beta(t) - \beta^*\}^2 dF(t) \approx \log \int \left\{ \frac{E(Z - E_{\beta^*}(Z|t)|t)^2}{V_{\beta^*}(Z|t)} \right\} dF(t) \tag{3.13}$$

may be worth investigating for this purpose. Under the proportional hazards model $\phi = 0$. Negative values for ϕ suggest overfitting and values greater than 1 ought correspond to departures from the proportional hazards assumption. In practice we would replace ϕ by the estimate $\hat{\phi}$ where

$$\hat{\phi} = \log \int \mathcal{V}_{\tilde{\beta}}^{-1}(Z|t) \{\mathcal{Z}(t) - \mathcal{E}_{\tilde{\beta}}(Z|t)\}^2 d\hat{F}(t) \tag{3.14}$$

More work is needed on this to obtain insight into what constitues a large positive or negative value. It is not clear at this time that, despite some intuitve appeal for the index, it measures what it is we would like it to measure. More study is needed before any recommendations could be given.

4 Testing for proportional against non proportional hazards

The literature on goodness of fit tests for the Cox model is quite vast and we do not propose any kind of a review here. A fairly extensive review of the most commonly used approaches can be found in O'Quigley and Xu (1998). One of the most

satisfactory tests, against non specific non proportional hazards alternatives, was the test developed by Wei (1984). Wei's test is attractive since it avoids arbitrary subdivisions of the time scale as suggested by other authors. The idea of Wei was to view the score statistic of equation 2.5, as a stochastic process, behaving like a Brownian bridge for large n. Specifically he considered the partial scores;

$$U_2(\beta, t) = \int_0^t \{\mathcal{Z}(s) - \mathcal{E}_\beta(Z|s)\} d\bar{N}(s) \tag{4.1}$$

and took as test statistic $\sup_t U_2(\hat{\beta}, t)$, large values indicating departures away from proportional hazards in the direction of non proportional hazards. Further work on this approach (Lin, Wei, and Ying, 1993) investigated more general processes than the score statistic, so that a wide choice of functions, potentially describing different kinds of departures from the model, are available. Rather than appeal to a large sample theory based on the Brownian bridge, in view of difficulties in the multivariate case, the authors developed a theory stemming from the martingale central limit theorem and the simulation of Gaussian processes.

Theorem 1 of this present paper can throw more light on the approach of Lin, Wei, and Ying (1993) and can, in particular, enable us to once again lean upon the large sample approximations based on the Brownian bridge. The increments of the process $\int_0^t \mathcal{Z}(s) d\bar{N}(s)$ at $t = X_i$ have mean $E(Z|X_i)$ and variance $V(Z|X_i)$. We can view these increments as being independent (Cox 1975; Andersen and Gill, 1982). Thus only the existence of the variance is necessary to be able to appeal to the functional central limit theorem. Standard results can then be applied, specifically those based on the supremum of a Brownian bridge over the interval (0,1).

To see this, consider the process $U^*(\beta, u)$, $(0 < u < 1)$, in which

$$U^*\left(\beta, \frac{j}{k}\right) = \frac{1}{\sqrt{k}} \int_0^{t_j} \mathcal{V}_\beta(Z|s)^{-1/2} dU_2(\beta, s)\,, \;\; j = 1, ..., k. \tag{4.2}$$

where $t_j = \bar{N}^{-1}(j)$. This process is only defined on k equispaced points of the interval (0,1] but we extend our definition to the whole interval via linear interpolation so that, for $u \in (\frac{j}{k}, \frac{j+1}{k})$ we write;

$$U^*(\beta, u) = U^*\left(\beta, \frac{j}{k}\right) + \{uk - j\} \left\{U^*\left(\beta, \frac{j+1}{k}\right) - U^*\left(\beta, \frac{j}{k}\right)\right\} \tag{4.3}$$

As n goes to infinity, under the usual Breslow and Crowley conditions, then we have that $U^*(\beta, u)$ converges in distribution to a Gaussian process with mean zero and variance equal to u. This is Brownian motion. Replacing β by a consistent estimate leaves asymptotic properties unaltered. For the purposes of inference we can consider the transformation

$$U_0^*(\hat{\beta}, u) = U^*(\hat{\beta}, u) - uU^*(\hat{\beta}, 1)$$

and note that $U_0^*(\hat{\beta}, u)$ converges in distribution to a Brownian bridge whereby;

$$\Pr\left\{\sup_u |U_0^*(\hat{\beta}, u)| \leq \alpha\right\} \rightarrow 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}\exp(-2k^2\alpha^2)\,, \quad \alpha \geq 0 \tag{4.4}$$

The multivariate case does not present any real additional complexity. Consider for example a two variable model as given by 2.11. Then we can extend the definition of $\mathcal{Z}(t)$ to two dimensions in an obvious way, i.e. a vector with components $\mathcal{Z}_1(t)$ and $\mathcal{Z}_2(t)$, step functions with discontinuities at the points X_i, $i = 1, ..., n$, where they take the values $Z_{1i}(X_i)$ and $Z_{2i}(X_i)$ respectively. For this two dimensional case we consider the increments in the process

$$\int_0^t \{\beta_1 \mathcal{Z}_1(s) + \beta_2 \mathcal{Z}_2(s)\}\, d\bar{N}(s)$$

at $t = X_i$, having mean

$$\beta_1 E(Z_1|X_i) + \beta_2 E(Z_2|X_i)$$

and variance

$$\beta_1^2 V(Z_1|X_i) + \beta_2^2 V(Z_2|X_i) + 2\beta_1\beta_2 \mathrm{Cov}\,(Z_1, Z_2|X_i).$$

The remaining steps now follow through just as in the one dimensional case, β_1 and β_2 being replaced by $\hat{\beta}_1$ and $\hat{\beta}_2$ respectively, and the conditional expectations, variances and covariances being replaced using formulae 2.13 and 2.14. It should now be clear how to deal with yet higher dimensions. It would also be possible to consider functions other than simple ones of the covariate Z. For time dependent Z we would need respect a requirement of predictability, in other words at time t we only use information available at all times strictly less than t, but, otherwise, great generality is possible.

5 Estimating survival given $Z \in H$

Although interest focusses mostly on estimation of the regression coefficients we are also interested in estimation of survival, especially survival conditional upon the covariable Z being restricted to some given subset H. When we wish to ignore any restrictions on Z as imposed by H then this is the marginal survival, most often estimated non parametrically via the Kaplan-Meier estimate. This appears to be a natural starting point to survival estimation, from which we can see how the Kaplan-Meier estimate is modified as we condition on $Z \in H$. In view of the earlier theorem 1, this turns out to be particularly simple to do.

A direct application of Bayes' formula gives,

$$S(t|Z \in H) = \frac{\int_t^\infty \mathrm{P}(Z \in H|t)dF(t)}{\int_0^\infty \mathrm{P}(Z \in H|t)dF(t)}. \tag{5.1}$$

Theorem 1 implies that $\mathrm{P}(Z \in H|t)$ can be consistently estimated by

$$\hat{P}(Z \in H|t) = \sum_{\{j:Z_\ell \in H\}} \pi_\ell(\beta(t), t) = \frac{\sum_H Y_\ell(t) \exp\{\hat{\beta}(t)Z_\ell\}}{\sum Y_\ell(t) \exp\{\hat{\beta}(t)Z_\ell\}}.$$

Again let $\hat{F}(t) = 1 - \hat{S}(t)$ be the Kaplan-Meier estimator of $F(t)$. Let $0 = t_0 < t_1 < ... < t_k$ be the distinct failure times, and let $W(t_i)$ still be the stepsize at t_i of the Kaplan-Meier curve as in Section 3. If the last observation is a failure, then

$$\hat{S}(t|Z \in H) = \frac{\int_t^\infty \hat{P}(Z \in H|t)d\hat{F}(t)}{\int_0^\infty \hat{P}(Z \in H|t)d\hat{F}(t)} = \frac{\sum_{t_i>t} \hat{P}(Z \in H|t_i)W(t_i)}{\sum_{i=1}^k \hat{P}(Z \in H|t_i)W(t_i)}. \tag{5.2}$$

Note also that conditioning further upon having already survived to time s, we obtain an equally simple expression

$$\hat{S}(t+s|Z \in H, T > s) = \frac{\int_{t+s}^\infty \hat{P}(Z \in H|u)d\hat{F}(u)}{\int_s^\infty \hat{P}(Z \in H|u)d\hat{F}(u)} = \frac{\sum_{t_i>t+s} \hat{P}(Z \in H|t_i)W(t_i)}{\sum_{t_i>s} \hat{P}(Z \in H|t_i)W(t_i)}. \tag{5.3}$$

When the last observation is not a failure and $\sum_1^k W(t_i) < 1$, an application of the law of total probability indicates that the following quantity should be added to both the numerator and the denominator in the above formula.

$$\hat{P}(Z \in H|T > t_k)\hat{S}(t_k). \tag{5.4}$$

In addition, using the empirical estimate over all the subjects that are censored after the last observed failure, we have

$$\hat{P}(Z \in H|T > t_k) = \frac{\sum_H Y_j(t_k+)}{\sum_1^n Y_j(t_k+)}, \tag{5.5}$$

where t_k+ denotes the moment right after time t_k. Therefore we can write

$$\hat{S}(t|Z \in H) = \frac{\sum_{t_i>t} \hat{P}(Z \in H|t_i)W(t_i) + \hat{P}(Z \in H|T > t_k)\{1 - \sum_1^k W(t_i)\}}{\sum_1^k \hat{P}(Z \in H|t_i)W(t_i) + \hat{P}(Z \in H|T > t_k)\{1 - \sum_1^k W(t_i)\}}. \tag{5.6}$$

The above estimate of the conditional survival function is readily calculated, since each term derives from standard procedures of survival analysis to fit the Cox model. In practice we use an estimate of the log-relative risk in the above computation. Note that when H includes all the possible values of z, $\hat{S}(t|Z \in H)$ simply becomes the Kaplan-Meier estimate of the marginal survival function.

Extension to the multivariate case takes place via the prognostic index. Rather than view H as some subspace of p dimensional space when we have p covariables, it makes more sense to transform everything to the real line via the linear combination $\sum \beta_i Z_i$. We can then consider different partitions of the real line and the consequences upon survival for groups of subjects having a prognostic index lying in any given interval.

References

[1] Altman, D.G. and Andersen, P.K. (1986): A note on the uncertainty of a survival probability estimated from Cox's regression model. *Biometrika*, **73**, 722-724.

[2] Andersen, P.K. (1991): Survival analysis 1982-1991: the second decade of the proportional hazards regression model. *Statistics in Medicine*, **10**, 1931-1941.

[3] Andersen, P.K., Christensen, E., Fauerholdt, L., and Schlichting, P. (1983): Measuring prognosis using the proportional hazards model. *Scand. J. Statist.*, **10**, 49-52.

[4] Andersen, P.K. and Gill, R.D. (1982): Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, **10**, 1100-1120.

[5] Anderson, G.L. and Fleming, T.R. (1995): Model misspecification in proportional hazards regression. *Biometrika*, **82**, 527-541.

[6] Bednarski, T. (1989): On sensitivity of Cox's estimator. *Statis. Decisions*, **7**, 215-228.

[7] Bednarski, T. (1993): Robust estimation in Cox's regression model. *Scand. J. Statist.*, **20**, 213-225.

[8] Breslow, N. (1972): Contribution to the discussion of paper by D.R. Cox. *J.R. Statist. Soc. B*, **34**, 216-217.

[9] Breslow, N. (1974): Covariance analysis of censored survival data. *Biometrics*, **30**, 89-99.

[10] Bretagnolle, J. and Huber-Carol, C. (1988): Effects of omitting covariates in Cox's model for survival data. *Scand. J. Statist.*, **15**, 125-138.

[11] Burr, D. (1994): A comparison of certain bootstrap confidence intervals in the Cox model. *JASA*, **89**, 1290-1302.

[12] Cox, D.R. (1972): Regression models and life tables (with discussion). *J.R. Statist. Soc. B*, **34**, 187-220.

[13] Cox, D.R. (1975): Partial likelihood. *Biometrika*, **62**, 269-276.

[14] Feller, W. (1966): *An Introduction to Probalility Theory and its Applications, Vol. II.* John Wiley & Sons, 2nd corrected printing.

[15] Ford, I., Norrie, J., and Ahmadi, S. (1995): Model inconsistency, illustrated by the Cox proportional hazards model. *Statistics in Medicine*, **14**, 735-746.

[16] Freireich, E.O. et al. (1963): The effect of 6-mercaptopmine on the duration of steroid induced remission in acute leukemia. *Blood*, **21**, 699-716.

[17] Gail, M.H., Wieand, S., and Piantadosi, S. (1984): Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, **71**, 431-444.

[18] Hougaard, P. (1986): Survival models for hetrogeneous populations derived from stable distributions. *Biometrika*, **73**, 387-396.

[19] Kalbfleisch, J.D. and Prentice, R.L. (1980): *The Statistical Analysis of Failure Time Data.* John Wiley and Sons.

[20] Kaplan, E.L. and Meier, P. (1958): Non-parametric estimation from incomplete observations. *JASA*, **53**, 457-481.

[21] Keiding, N. (1995): Historical controls and modern survival analysis. *Lifetime Data Analysis*, **1**, 19-25.

[22] Keiding, N. and Knuiman, M.W. (1990): Letter to the editor: survival analysis in natural history studies of disease. *Statistics in Medicine*, **9**, 1221-1222.

[23] Klein, J.P., Lee, S.C., and Moeschberger, M.L. (1990): A partially parametric estimator of survival in the presence of randomly censored data. *Biometrics*, **46**, 795-811.

[24] Lagakos, S.W. and Schoenfeld, D.A. (1984): Properties of proportional-hazards score tests under misspecified regression models. *Biometrics*, **40**, 1037-1048.

[25] Lancaster, T. and Nickell, S. (1980): The analysis of re-employment probabilities for the unemployed. *J.R. Statist. Soc. A*, **143**, 141-165.

[26] Liang, K.Y. and Zeger, S.L. (1986): Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

[27] Liang, K.Y. and Zeger, S.L. (1995): Inference based on estimating functions in the presence of nuisance parameters. *Stat. Science*, **10**, 158-173.

[28] Lin, D.Y. (1991): Goodness-of -fit analysis for the Cox regression model based on a class of parameter estimators. *JASA*, **86**, 725-728.

[29] Lin, D.Y. and Wei, L.J. (1989): The robust inference for the Cox proportional hazards model. *JASA*, **84**, 1074-1078.

[30] Lin, D.Y., Fleming, T.R., and Wei, L.J. (1994): Confidence bands for survival curves under the proportional hazards model. *Biometrika*, **81**, 73-81.

[31] Link, C.L. (1984): Confidence intervals for the survival function using Cox's proportional-hazard model with covariates. *Biometrics*, **40**, 601-610.

[32] Moeschberger, M.L. and Klein, J.P. (1985): A comparison of several methods of estimating the survival function when there is extreme right censoring. *Biometrics*, **41**, 253-259.

[33] Newton, M.A. and Raftery, A.E. (1994): Approximate Bayesian inference with the weighted likelihood bootstrap. *J.R. Statist. Soc. B*, **56**, 3-26.

[34] Oakes, D. (1986): An approximate likelihood procedure for censored data. *Biometrics*, **42**, 177-182.

[35] O'Quigley, J. and Pessione, F. (1989): Score tests for homogeneity of regression effect in the proportional hazards model. *Biometrics*, **45**, 135-144.

[36] O'Quigley, J. and Pessione, F. (1991): The problem of a covariate-time qualitative interaction in a survival study. *Biometrics*, **47**, 101-115.

[37] Rashid, S.A., O'Quigley, J., Cooper, E.H., and Giles, G. (1982): Plasma protein profiles and prognosis in gastric cancer. *Brit. J. Cancer*, **45**, 390-394.

[38] Sasieni, P. (1993): Maximum weighted partial likelihood estimators for the Cox model. *JASA*, **88**, 144-152.

[39] Schoenfeld, D.A. (1980): Chi-squared goodness-of-fit tests for proportional hazards regression model. *Biometrika*, **67**, 145-53.

[40] Struthers, C.A. and Kalbfleisch, J.D. (1986): Misspecified proportional hazard models. *Biometrika*, **73**, 363-369.

[41] Stute, W. (1995): The central limit theorem under random censorship. *The Annals of Statistics*, **23**, 422-439.

[42] Tsiatis, A.A. (1975): A nonidentifiability aspect of the problem of competing risks. *Proc. Nat. Acad. Sci. USA*, **72**, 20-22.

[43] Tsiatis, A.A. (1981): A large sample study of Cox's regression model. *The Annals of Statistics*, **9**, 93-108.

[44] Verweij, J.A. and van Houwelingen, H.A. (1995): Time-dependent effects of fixed covariates in Cox regression. *Biometrics*, **51**, 1550-1556.

[45] Xu, R. (1996): Inference for the proportional hazards model. *Ph.D. thesis of University of California, San Diego.*