Cox Regression Model

The Cox or **proportional hazards** regression model [21] is used to analyze **survival** or failure time data. It is now perhaps the most widely used statistical model in medical research. Whenever the outcome of a **clinical trial** is the time to an event, the Cox model is the first method considered by most researchers. The model has also inspired an enormous statistical literature, ranging from the mathematical study of estimating the model parameters, to applied techniques for validating the model assumptions.

This article is divided into sections touching on some of the vast literature that has developed around the model:

1. model definition
2. history
3. using the Cox model–the basics
4. estimators and **algorithms**
5. asymptotic properties (*see* **Large-sample Theory**)
6. **time-dependent** explanatory variables
7. **model checking**
8. alternatives and extensions.

Several books have now been published on survival analysis that devote major sections to the Cox model. The first of these appeared in the early 1980s [23, 50]. Of the more recent books some are mathematically rigorous [6, 29], while others are more applied [20, 53, 60]. The book by Andersen et al. [6] is the most comprehensive.

Model Definition

Cox's essential novelty was to model the hazard function (*see* **Hazard Rate**) rather than the mean or some other measure of location. Let X denote a random failure time and $\mathbf{Z}$ a vector of **explanatory variables**. The conditional hazard of X given $\mathbf{Z} = \mathbf{z}$ at time t is defined as

$$\lambda(t|\mathbf{z}) = \lim_{\Delta t \downarrow 0} \frac{\Pr(X \leq t + \Delta t | X > t, \mathbf{z})}{\Delta t}. \quad (1)$$

The hazard function is sometimes called the intensity function or the force of mortality. Roughly, the hazard function is the probability that someone who is alive now will die in the next small unit of time. Cox proposed that the conditional hazard be modeled as the product of an arbitrary baseline hazard $\lambda_0(t)$ and an exponential form that is linear in $\mathbf{z}$:

$$\lambda(t|\mathbf{z}) = \lambda_0(t) \exp(\boldsymbol{\beta}'\mathbf{z}). \quad (2)$$

Here $\boldsymbol{\beta}$ is a vector of regression parameters and the infinite-dimensional parameter $\lambda_0(\cdot)$ is the hazard function for an individual with $\mathbf{Z} = \mathbf{0}$. The model in (2) forces the hazard ratio between two individuals to be constant over time:

$$\frac{\lambda(t|\mathbf{z}_2)}{\lambda(t|\mathbf{z}_1)} = \exp[\boldsymbol{\beta}'(\mathbf{z}_2 - \mathbf{z}_1)].$$

The exponential form of the relative risk function has become standard and is the most stable computationally, but it is not the only possibility. The more general model,

$$\lambda(t|\mathbf{z}) = \lambda_0(t) r(\boldsymbol{\beta}'\mathbf{z}),$$

for some known function r has also been considered [67, 82].

History

A distinguishing feature of survival data is that it is subject to **censoring**. Very often one does not observe the survival time for all individuals in a study. One may only know that a certain individual was still alive at some time T^*. If T_i^* is the last time at which individual i is known to be alive, it is called a censoring time – the individual's follow-up was censored at T_i^*. In 1958, Kaplan & Meier [51] studied the product-limit estimator of a survival function based on censored data (*see* **Kaplan–Meier Estimator**). The key concept of viewing the data as a process that reveals itself over time can be seen in their paper. Test statistics for censored data were considered a few years later [31, 59], and some may view the Cox model as the natural generalization to a regression setting of ideas present in Mantel's writing [59]. At about the same time, Feigl & Zelen [28] considered various **exponential** regression models. One of their models is equivalent to the Cox model with the baseline hazard constrained to be constant for all time, so that $\lambda(t|\mathbf{z})$ is a function of $\mathbf{z}$ but not t. However, unlike Cox [21], they formulate the model in terms of a parameterization of the mean

survival time, even though they use the exponential assumption to predict the entire survival distribution.

Cox's 1972 paper [21] was instantly acclaimed as a breakthrough in the analysis of right censored data, as can been seen from the enthusiastic discussion published together with the article. The model was rapidly adopted by applied statisticians, particularly in clinical trials. Its use became widespread once user-friendly software became readily available. Today, one can hardly open a leading medical or statistical journal without finding at least one reference to Cox (1972)! It is one of the most widely cited papers in scientific literature.

The original paper introduced a model that was to revolutionize the field, and provided the estimator that is today programmed into many statistical software packages. There were, however, several issues that were to challenge the statistical community. Some of these, such as how to deal with ties (two or more individuals with the same failure time) [63] (*see* **Tied Survival Times**), and the basis for the proposed estimator, were addressed at the Royal Statistical Society meeting. Cox provided justification for the estimator himself by introducing the concept of a **partial likelihood** [22]. But it was not until later that the estimators were shown to be **efficient** [11, 27]. Formal proofs of **consistency** and asymptotic normality took nearly a decade [5, 83]. Another topic of considerable interest to statisticians is the effect of **misspecification** on the estimates [80], and model interpretation. Various types of misspecification have been considered: explanatory variables measured with error [65] (*see* **Errors in Variables**); omission of important explanatory variables [16, 54, 78]; and rare but gross data contamination [9, 72] (*see* **Outliers**).

Parallel with the theoretical progress was work on model building and model checking. The results were less satisfactory than the elegant theory that developed around **counting processes** and martingales, but a variety of tools are now available. These included **goodness of fit** tests, as well as **residuals** and other **diagnostics**. Andersen [4] and others have discussed the quality of presentation of Cox regression analyses in the medical literature. Despite their constructive suggestions, the "Methods" sections of many papers are still no more informative than "we used the Cox model".

The basic model, (2), has been generalized in various directions. Even the original paper [21] considered time-dependent covariates, but these still cause a variety of difficulties [3]. A simple generalization is to permit different baseline hazard functions in each of a number of strata (*see* **Stratification**). The stratified Cox model assumes that, within each stratum, the proportional hazards assumption is justified and that the effect of the variable $\mathbf{Z}$ is the same in all strata:

$$\lambda_j(t|\mathbf{z}) := \lambda(t|\mathbf{z},\ \text{stratum}\ j) = \lambda_{0j}(t)\exp(\boldsymbol{\beta}'\mathbf{z}). \quad (3)$$

By incorporating constructed variables, that are constant in some strata, the stratified model, (3), can be used to model interactions between explanatory variables and strata. Suppose, for example, that one is stratifying by sex and including age as an explanatory variable. Let $z_1 = (\text{age} - 50)$ for men, $= 0$ for women; and let $z_2 = (\text{age} - 50)$ for women, $= 0$ for men. Then a model stratified on sex that includes z_1, z_2, and a treatment indicator z_3 permits interactions between age and sex, but assumes that the treatment acts proportionately on the hazards for any age–sex combination.

Many models used for analysis of **multivariate survival data** are generalizations of the Cox model, but they are not discussed here.

Using the Cox Model – the Basics

Before using the Cox model, or even attempting to interpret a published analysis, one must have some understanding of the assumptions that underlie the analysis. This section discusses those assumptions and explains a typical output from fitting the model in a statistical package.

There are three components to the data on each individual: the possibly censored failure time T; an indicator δ (*see* **Dummy Variables**) equal to 1 if T is a true failure time, 0 if it is censored; and $\mathbf{Z}$, the vector of explanatory variables. The model is flexible enough to incorporate explanatory variables that change value over the course of the study, but in this section we assume that $\mathbf{Z}$ is fixed and measured at time $t = 0$. The key censoring assumption is that the observation $(T = t, \delta = 0)$ tells us nothing more than that the true failure time X is greater than t.

In a clinical trial, the time origin for each individual will usually be his or her time of entry into the trial. If the trial ends at a particular calendar time, censoring all individuals who are not yet dead, then the censoring times are the times from entry until

the end of the trial and will vary from one individual to another. This is called administrative (or progressive type I) censoring. In such situations, it is necessary for survival to be independent of entry time for the above condition to be satisfied. To some extent this can be examined by including entry time as a covariate or by stratifying on the date of entry. Other forms of censoring are more problematic. If, for instance, a patient emigrates, one needs to consider whether this implies that the patient had in fact recovered. Conversely, a patient who fails to attend a follow-up clinic might be too sick to get out of bed. In such cases, the fact that the patient was censored at t tells us rather more than that she was alive at t.

The Cox model itself makes three assumptions: first, that the ratio of the hazards of two individuals is the same at all times; secondly, that the explanatory variables act multiplicatively on the hazard; and thirdly, that, conditionally on $\mathbf{Z}_i$ and $\mathbf{Z}_j$, the failure times of individuals i and j are independent. As with all regression models, one also assumes that the explanatory variables have been transformed so that they may be entered without further transformation and that all interactions have been included explicitly. We will see in the section on asymptotics that the independence assumption can be relaxed.

Table 1 presents the results of fitting a Cox model to data from 216 patients with primary biliary cirrhosis in a clinical trial of azathioprine vs. placebo [18]. The six variables were selected from an initial set of 25 partly using forward stepwise selection. An additional 32 patients were excluded because they had missing values of one or more of the six variables. Recruitment was over 6 years and follow-up a further 6 years. Of the 216 patients, 113 had censored survival times. The regression coefficients may be combined with their standard errors to obtain **confidence intervals** that rely on the asymptotic normality of the estimates.

The positive coefficient associated with treatment implies that patients on the placebo ($Z = 1$) had poorer prognosis than those on azathioprine ($Z = 0$): the hazard of those on placebo is about 1.7 times greater than that of those on active treatment. Similarly, older patients had poorer prognosis. The hazard ratio associated with two patients aged 50 and 30 is $\exp[0.0069(\exp 3 - \exp 1)] = 1.13$. Notice, however, that the effect on survival is not fully described by the information in Table 1 because, without estimating the baseline hazard, one cannot translate the regression coefficients into effects on 5-years survival nor on **median survival**.

Most statistical software for Cox regression will also estimate the cumulative baseline hazard function

$$\Lambda_0(t) = \int_0^t \lambda_0(u)\,\mathrm{d}u \tag{4}$$

(*See* **Survival Distributions and Their Characteristics**), and from this one can calculate the estimated survival function for a given $\mathbf{z}$:

$$\Pr(X > t|\mathbf{z}) = \prod_{\{i:T_i \le t\}} [1 - \mathrm{d}\hat{\Lambda}_0(T_i)\exp(\boldsymbol{\beta}'\mathbf{z})].$$

Plots of the estimated survival function can be made for various $\mathbf{z}$s, and these can be viewed like **Kaplan–Meier** graphs. Alternatively, the estimated survival function can be used to estimate 5-year survival, say, as a function of the prognostic index $\boldsymbol{\beta}'\mathbf{z}$ (*see* **Prognosis**).

Estimators and Algorithms

The regression coefficients $\boldsymbol{\beta}$ are estimated by maximizing the so-called partial likelihood $L(\boldsymbol{\beta})$ [22]. An

Table 1 Cox model fitted to data from a clinical trial comparing the effects of azathioprine and placebo on the survival of 216 patients with primary biliary cirrhosis [18]. The six variables shown were selected, partly by a forward stepwise procedure, from 25 candidate variables

Variable	Coding	Coeff. $\hat{\beta}$	se($\hat{\beta}$)	exp($\hat{\beta}$)
Serum bilirubin	$\log_{10}$ (value in μmol/l)	2.51	0.316	12.3
Age	exp[(age in years − 20)/10]	0.0069	0.0016	1.0
Cirrhosis	0 = No; 1 = Yes	0.88	0.216	2.4
Serum albumin	value in g/l	−0.0504	0.018	0.95
Central cholestasis	0 = No; 1 = Yes	0.68	0.275	2.0
Therapy	0 = azathioprine; 1 = placebo	0.52	0.201	1.7

individual is said to be at risk at t if he has not yet failed nor been censored. This concept can be generalized to allow for individuals who do not enter the study at time 0. Such **delayed entry**, or left **truncation**, as it is called, often arises when t is the age of a patient or the time from infection, so that patients enter the study at some time $T_i^0 > 0$. Consider $L_i(\boldsymbol{\beta})$, the conditional probability that individual i fails at time T_i given that exactly one individual fails at T_i and knowing the values of $\mathbf{Z}$ for all individuals at risk at T_i:

$$L_i(\boldsymbol{\beta}) = \frac{\lambda(T_i|\mathbf{Z}_i)}{\sum_{j\in R_i}\lambda(T_i|\mathbf{Z}_j)} = \frac{\exp(\boldsymbol{\beta}' Z_i)}{\sum_{j\in R_i}\exp(\boldsymbol{\beta}'\mathbf{Z}_j)}, \qquad (5)$$

where $R_i = \{j : T_j^0 < T_i \le T_j\}$ is the **risk set** just prior to T_i. The partial likelihood is the product of these **conditional probabilities** over all failure times: $L(\boldsymbol{\beta}) = \prod_i L_i(\boldsymbol{\beta})$. Notice that the partial likelihood is a function of $\boldsymbol{\beta}$ only – it does not depend on the baseline hazard $\lambda_0(\cdot)$. With certain types of censoring (or no censoring) the partial likelihood is just the marginal **likelihood** of the ranks of the failure times. If there are ties in the data (two or more individuals failing at the same time), then both the partial likelihood and the marginal likelihood become difficult computationally [50, pp. 74–78]. Instead, most packages use an approximation [13, 63]:

$$L_i(\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}'\mathbf{S}_i)}{\left[\sum_{j\in R_i}\exp(\boldsymbol{\beta}'\mathbf{Z}_j)\right]^{d_i}}, \qquad (6)$$

where d_i is the number of individuals failing at T_i and $\mathbf{S}_i$ is the sum of the $\mathbf{Z}_j$ for these d_i individuals. The approximation is reasonable provided the number of ties at any failure time is small compared to the number in the risk set. Note that i indexes the N distinct failure times, whereas j indexes the n individuals ($n \ge N$).

It is standard practice to maximize the partial likelihood using Newton–Raphson to find a $\boldsymbol{\beta}$ at which the derivative of its logarithm is zero (*see* **Optimization and Nonlinear Equations**). Indeed, Jacobsen [47] has shown that, when the relative risk function $r(\boldsymbol{\beta}'\mathbf{z}) = \exp(\boldsymbol{\beta}'\mathbf{z})$, $l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta})$ is concave. (It is strictly concave provided there is no exact **collinearity** among the explanatory variables and that no linear combination of the variables is a perfect predictor of failure. The latter would imply an infinite observed hazard ratio.)

We use the following notation: let

$$\mathbf{S}^{(k)}(\boldsymbol{\beta}, T_i) = \sum_{j\in R_i}\mathbf{Z}^{\otimes k}\exp(\boldsymbol{\beta}'\mathbf{Z}_j),$$

where $\mathbf{Z}^{\otimes 0} = 1$, $\mathbf{Z}^{\otimes 1} = \mathbf{Z}$, and $\mathbf{Z}^{\otimes 2} = \mathbf{Z}\mathbf{Z}'$. Let $\mathbf{U}(\boldsymbol{\beta})$ denote the score

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_i \frac{\mathrm{d}\log L_i(\boldsymbol{\beta})}{\mathrm{d}\boldsymbol{\beta}} = \sum_i \left[\mathbf{S}_i - d_i\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, T_i)}{\mathbf{S}^{(0)}(\boldsymbol{\beta}, T_i)}\right], \qquad (7)$$

and $\mathbf{I}(\boldsymbol{\beta})$ minus the Hessian:

$$\mathbf{I}(\boldsymbol{\beta}) = -\frac{\mathrm{d}\boldsymbol{U}(\boldsymbol{\beta})}{\mathrm{d}\boldsymbol{\beta}} = \sum_i d_i \times \left\{\frac{\mathbf{S}^{(2)}(\boldsymbol{\beta}, T_i)}{\mathbf{S}^{(0)}(\boldsymbol{\beta}, T_i)} - \left[\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, T_i)}{\mathbf{S}^{(0)}(\boldsymbol{\beta}, T_i)}\right]^{\otimes 2}\right\}. \qquad (8)$$

Given an estimate $\boldsymbol{\beta}^{(m)}$, one step of the algorithm gives

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \mathbf{I}(\boldsymbol{\beta}^{(m)})^{-1}\boldsymbol{U}(\boldsymbol{\beta}^{(m)}).$$

The algorithm is generally started from $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ and convergence is determined by the magnitude of $|\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}|$.

When there are S strata, one considers those at risk in each stratum separately. Let R_{si} denote the set of indices of individuals in stratum s at risk at time T_i, and let $L_{si}(\boldsymbol{\beta})$ be the partial likelihood contribution from stratum s and time T_i. Note that $\mathbf{S}_{si}$ is the sum of the $\mathbf{Z}_j$ of the d_{si} individuals in stratum s who fail at time T_i. The partial likelihood is then simply the product of the stratum specific partial likelihoods:

$$L(\boldsymbol{\beta}) = \prod_{s=1}^{S}\prod_i L_{si}(\boldsymbol{\beta}).$$

Although the partial likelihood is not in general a likelihood, it is usually treated as such. It is standard practice to report the value of the logarithm of the partial likelihood and to compare the partial likelihood ratio statistic to a **chi-square distribution** for testing between nested regression models (*see* **Likelihood Ratio Tests**). Similarly, the covariance of $\hat{\boldsymbol{\beta}}$

is estimated by $\mathbf{I}(\hat{\boldsymbol{\beta}})^{-1}$ and score tests (*see* **Likelihood**) are based on $\mathbf{U}(\mathbf{0})'\mathbf{I}(\mathbf{0})^{-1}\mathbf{U}(\mathbf{0})$. Indeed, in the absence of ties ($d_{si} = 1$ for all s and i), the score test from the Cox model with $K - 1$ dummy variables corresponding to a factor with K levels is identical to the K-sample log rank test. Further, the stratified log rank test is identical to the score test from the stratified Cox model.

Having computed $\hat{\boldsymbol{\beta}}$, the estimated regression coefficients, one can calculate the Breslow estimate of the cumulative baseline hazard [13] explicitly. The estimator for stratum s is

$$\hat{\Lambda}_{s0}(t) = \sum_{i:T_i \le t} \frac{d_{si}}{\sum_{j \in R_{si}} \exp(\hat{\boldsymbol{\beta}}'\mathbf{Z}_j)}. \tag{9}$$

Estimation of the hazard function itself can be done by taking a smooth derivative of the cumulative hazard. This is usually achieved by the kernel method [68] (*see* **Density Estimation**). The jumps in the Breslow estimate should not be used without **smoothing**. The jump at T_i crudely approximates $\lambda_0(T_i)(T_i - T_{i-1})$ not $\lambda_0(T_i)$. Breslow [13] also showed that the maximum partial likelihood estimate of $\boldsymbol{\beta}$ and the estimated cumulative baseline hazard, (5), can also be obtained by maximizing the full likelihood for $\boldsymbol{\beta}$ and Λ_0 simultaneously, assuming that Λ_0 is piecewise linear **spline**, i.e. the hazard $\lambda_0(t)$ is constant between each pair of ordered failure times. This heuristic argument was made precise by Johansen [48]. He showed that, in certain circumstances, the partial likelihood is formally the profile likelihood for $\boldsymbol{\beta}$. He permitted Λ_0 to be a step function and assumed that at the jumps $\mathrm{d}\Lambda(t|\mathbf{z}) = \exp(\boldsymbol{\beta}'\mathbf{z})\,\mathrm{d}\Lambda_0(t)$.

During the 1970s anyone wishing to fit a Cox model had to use a stand-alone computer program such as the FORTRAN code provided in the book by Kalbfleisch & Prentice [50]. Today, however, the situation is very different and there are many commercially available general statistical packages that will fit a Cox model to large data sets (*see* **Software, Biostatistical**).

Asymptotic Properties

The large sample properties of the maximum partial likelihood estimator of $\boldsymbol{\beta}$ and of the Breslow estimator of Λ_0 are unsurprising, but proofs of these results took some time. When the Cox model holds with parameters $\boldsymbol{\beta}_0$ (and Λ_0), the distribution of $\hat{\boldsymbol{\beta}}$ can be approximated by **multivariate normal** with mean $\boldsymbol{\beta}_0$ and a **covariance matrix** that can be estimated by $\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$.

Two quite different approaches were successful. The first due to Tsiatis [83] was to consider independent and identically distributed triples $(X_i, \mathbf{Z}_i, C_i)$, where X_i is the failure time and C_i is the censoring time. It is assumed that the X_i are generated from a Cox model with covariates $\mathbf{Z}_i$ and that X_i are conditionally independent of C_i given $\mathbf{Z}_i$. The observed data are $(T_i, \mathbf{Z}_i, D_i)$, $i = 1, \ldots, n$, where $T_i = \min(X_i, C_i)$ and $D_i = 1$ if $T_i = X_i$ (the event is observed), and $D_i = 0$ otherwise (the event is censored). The estimators are functionals of the observed data, and classical large sample theory is applied. Under this model it can be shown that

$$\frac{\mathbf{S}^{(1)}(\hat{\boldsymbol{\beta}}, t)}{\mathbf{S}^{(0)}(\hat{\boldsymbol{\beta}}, t)} \to \mathbf{E}(\mathbf{Z}|T = t, D = 1)$$

and that $\mathbf{I}(\hat{\boldsymbol{\beta}})/n \to \mathbf{E}[D\mathrm{var}(\mathbf{Z}|T, D)]$ [70]. By viewing the estimators as functionals of the empirical distribution of the unobserved triples and using results from the theory of empirical processes, it is possible to study the large sample properties of the Cox estimators even when the data come from some other model [72].

The other approach to large sample theory using a martingale **central limit theory** requires reformulating the model. This approach adds much insight to the model and will be outlined here. The counting process view of survival analysis is due to Aalen [1]. Andersen & Gill [5] redefined the Cox model and provided elegant proofs of its large-sample properties under mild regularity conditions.

Counting Process Formulation

A multivariate counting process

$$N = \{N_i(t) : 0 \le t < \infty; i = 1, \ldots, n\}$$

is a nondecreasing integer-valued stochastic process with n components. It is assumed that $N_i(0) = 0$ for all i and that the jumps are all of size +1. The process may count the number of events that have occurred in each of n individuals by time t. If the event is the death of a person, then $N_i(t) \in \{0, 1\}$ since people only die once! For technical reasons, N_i is taken to be

right continuous (so that $N_i(t)$ represents the number of events in $[0, t]$) and no two components of N jump at the same time.

Associated with such a counting process is a cumulative intensity process A with components defined by

$$A_i(t + \mathrm{d}t) - A_i(t) = \Pr\{N_i(t + \mathrm{d}t) - N_i(t) = 1 | \mathcal{F}_{t-}\},$$

where $\mathcal{F}_{t-}$ represents everything that has happened until just before t. The history $\mathcal{F}_{t-}$ will certainly include the paths of $N_j(\cdot)$ on $[0, t)$, $j = 1, \ldots, n$, and may include other information such as censoring or explanatory variables from $[0, t)$. $M = N - A$ is a multivariate martingale with respect to the history (filtration) $\{\mathcal{F}_t : t \geq 0\}$. The Andersen & Gill [7] generalization of the Cox model is that

$$A_i(t + \mathrm{d}t) - A_i(t) = \alpha_i(t)\,\mathrm{d}t = Y_i(t)\lambda_0(t) \times \exp[\boldsymbol{\beta}_0' \mathbf{Z}_i(t)]\,\mathrm{d}t,$$

where $Y_i(t)$ is equal to 1 if individual i is under observation just before time t, and is equal to 0 otherwise. $Y_i(\cdot)$ is called the ith "at-risk" indicator process. Here we are assuming that the process A is absolutely continuous with derivative α. Note that we have written the explanatory variables as processes depending on t, and that the definition of the intensity process requires $\{\mathbf{Z}_i(u) : 0 \leq u \leq t, i = 1, \ldots, n\}$ to be in the history $\mathcal{F}_{t-}$. This means that the value of $\mathbf{Z}(t)$ should be known just before t.

The classical Cox model corresponds to a very simple counting process, each component of which jumps at most once. We have

$$N_i(t) = I\{T_i \leq t, T_i \leq C_i\}$$

and

$$Y_i(t) = I\{X_i \geq t, C_i \geq t\} = I\{T_i \geq t\}.$$

N_i starts at 0 and jumps to one when individual i is observed to die. If individual i is censored, N_i remains 0 for ever. Recall that $\alpha_i(t)\,\mathrm{d}t$ is the probability of N_i jumping in the interval $[t, t + \mathrm{d}t]$. If individual i has died or been censored before time t, then there is no chance of observing a death in the interval $[t, t + \mathrm{d}t]$, so $\alpha_i(t) = 0$. Otherwise $\alpha_i(t) = \lambda(t|\mathbf{Z}_i)$ by the definition of the hazard function. Hence in general $\alpha_i(t) = Y_i(t)\lambda(t|\mathbf{Z}_i)$.

Using the new notation, we define the log partial likelihood using information up to time u as

$$l(\boldsymbol{\beta}, u) = \int_0^u \sum_{i=1}^n \left(\boldsymbol{\beta}' \mathbf{Z}_i(t)\,\mathrm{d}N_i(t) - \log\left\{ \sum_{j=1}^n Y_j(t) \exp[\boldsymbol{\beta}' \mathbf{Z}_j(t)] \right\} \mathrm{d}N_i(t) \right).$$

Note that $\mathrm{d}N_i(t)$ is equal to either 0 or 1, because N_i is a counting process. Thus integration with respect to $\mathrm{d}N_i(t)$ is simple: in the classical Cox model $\int f(t)\,\mathrm{d}N_i(t) = D_i f(T_i)$. Differentiate l with respect to $\boldsymbol{\beta}$ to get the score process

$$\mathbf{U}(\boldsymbol{\beta}, u) = \int_0^u \sum_{i=1}^n [\mathbf{Z}_i(t) - \mathbf{E}(\boldsymbol{\beta}, t)]\,\mathrm{d}N_i(t),$$

where

$$\mathbf{E}(\boldsymbol{\beta}, t) = \frac{\sum_{j=1}^n Y_j(t) \mathbf{Z}_j(t) \exp[\boldsymbol{\beta}' \mathbf{Z}_j(t)]}{\sum_{j=1}^n Y_j(t) \exp[\boldsymbol{\beta}' \mathbf{Z}_j(t)]}. \quad (10)$$

It is easy to show that at the true $\boldsymbol{\beta}$, integration with respect to the intensity process is identically zero (for all u). Hence, at $\boldsymbol{\beta}_0$, one may replace $\mathrm{d}N_i(t)$ by $\mathrm{d}M_i(t)$:

$$\mathbf{U}(\boldsymbol{\beta}_0, t) = \int_0^u \sum_{i=1}^n [\mathbf{Z}_i(t) - \mathbf{E}(\boldsymbol{\beta}_0, t)]\,\mathrm{d}M_i(t).$$

It follows from the theory of martingale transforms that $\mathbf{U}(\boldsymbol{\beta}_0, \cdot)$ is a martingale since the integrand $[\mathbf{Z}_i(t) - \mathbf{E}(\boldsymbol{\beta}_0, t)]$ is predictable (i.e. its value is known just prior to t). Under mild regularity conditions [5] one can apply a martingale central limit theorem to show that $n^{-1/2}\mathbf{U}(\boldsymbol{\beta}_0, \cdot)$ converges in distribution to a Gaussian process.

Extending the counting process notation in the obvious way to permit strata, so that, for instance, $Y_{si}(u)$ indicates whether individual i is at risk in

stratum s at time u, the Breslow estimator is

$$\hat{\Lambda}_{s0}(t) = \int_0^t \frac{\sum_{i=1}^{n} \mathrm{d}N_{si}(u)}{\sum_{i=1}^{n} Y_{si}(u) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_i(u)]}$$

$$= \int_0^t \sum_{i=1}^{n} \mathrm{d}N_{si}(u) / S_s^{(0)}(\hat{\boldsymbol{\beta}}, u)$$

Let $J_s(t) = I\{\sum_{i=1}^{n} Y_{si}(t)\} > 0$. Then

$$\int_0^t J_s(u)\, \mathrm{d}[\hat{\Lambda}_{s0}(u) - \Lambda_{s0}(u)] = \int_0^t \frac{J_s(u)}{S_s^{(0)}(\hat{\boldsymbol{\beta}}, u)}$$

$$\times \sum_{i=1}^{n} \{\mathrm{d}N_{si}(u) - Y_{si}(u) \exp[\boldsymbol{\beta}_0' \mathbf{Z}_i(u)]\, \mathrm{d}\Lambda_{s0}(u)\}$$

$$= \int_0^t \frac{J_s(u)}{S_s^{(0)}(\hat{\boldsymbol{\beta}}, u)} \sum_{i=1}^{n} \mathrm{d}M_{si}(u).$$

Thus, once again, the asymptotics can be proved using a martingale central limit theorem.

Time-Dependent Explanatory Variables

The possibility of including explanatory variables that change with time was realized by Cox in his original article [21]. There it is suggested that the inclusion of a user-defined variable $Z_2(t) = tZ_1$ might be used as a test of the proportional hazards assumption. Other authors have included explanatory variables that change value at possibly random times. The classical example of this sort of covariate is one that indicates whether a patient has received a heart transplant before time t [25]. The uses and interpretations of these two types of time-dependent variables are quite different. In this section they will be discussed relying heavily on the ideas presented by Kalbfleisch & Prentice [50].

External or Ancillary Variables

An external variable is one that is not affected by the failure process. The simplest sort of external variable is a fixed or time-independent one. A second type is a defined variable such as $Z_2(t) = tZ_1$. Although Z_2 is not fixed, its entire path is known from the outset. A more general example of an external variable is a measure of air pollution as a predictor of severe asthma attacks. Although the level of air pollution is not known in advance, it is "external" to the individuals in the study. Furthermore, the marginal distribution of the variable does not involve the parameters of the failure time model. The whole history of an external variable can be included in $\mathcal{F}_0$ and the hazard or intensity process can be related to the survival function $\Pr(T \geq t|\mathcal{F}_0)$ in the usual way.

Internal Variables

An internal explanatory variable is the output of a **stochastic process** that is generated by the individual under study and so is observed only so long as the individual survives and is uncensored [50]. An example might be the level of β-2 microglobulin in a patient's sera. In practice, the actual level at any given time will be unknown. Instead one uses the level as measured in the most recent blood sample. Typically blood will be taken at most a dozen times during a trial. In such circumstances, the term "updated" may be preferred to "time-dependent".

The key point is that although one may include the history of an internal process up to time t in the filtration $\mathcal{F}_t$ and so define the hazard or intensity function, the intensity function is itself a random process and is not simply a function of the survival function. In general survival from u to t depends on $\{Z(s) : u \leq s \leq t\}$ and this is unknown at u. Furthermore, if Z is only observed when an individual is alive, then $\Pr(T \geq t|Z(t)$ is not missing$) = 1$. Thus it is not possible to make predictions of survival from models that include internal explanatory variables. To do that one must jointly model the survival process and the explanatory variable trajectory.

In a clinical trial with primary focus on a treatment which is fixed by randomization at time 0, internal variables may change in response to treatment. If the effect of treatment is predominantly reflected in the changing value of the explanatory variable, a Cox model of survival that includes both treatment and the updated measurements of the explanatory variable will show little or no treatment differences. Clearly, then, one must be very careful when interpreting the output of a Cox model that includes an internal explanatory variable. Treatment differences in a

model that includes the values of explanatory variables only at time 0 may be inferred to be causative (because of **randomization**). When a large treatment difference is attenuated by inclusion of updated measurements of an internal variable, one may gain useful insights into the mechanism through which the treatment is effective. In such circumstances, it is sensible to also explore the effect of treatment on the internal variable directly.

As with censoring, the value of a variable may depend on the history of the trial so far, without depending on the history of a given individual. Thus, for instance, one might decide to change the environment of a controlled experiment after every 15 deaths. Such a variable is neither internal nor external, but for the purpose of making inference it is closer to an external process.

Computing with Time-Dependent Variables

There are many practical issues in fitting models using time-varying regressors, such as how to deal with missing values, that are not discussed here [3].

The Cox model does not distinguish between a single individual who enters a trial at time 0 and dies at time T_i with fixed regressors $\mathbf{Z}_i$, from two individuals both with regressors $\mathbf{Z}_i$ one of whom enters at time 0 and is censored at time u and one of whom enters at u and dies at T_i. This may sound surprising, but it is true; the likelihood contributions from $[0, u]$ and $(u, T_i]$ are $\Pr(X > u|\mathbf{Z}_i)$, and

$$\frac{\Pr(T_i \le X < T_i + \mathrm{d}t | X > u, \mathbf{Z}_i)}{\mathrm{d}t} = \frac{\Pr(T_i \le X < T_i + \mathrm{d}t|\mathbf{Z}_i)/\mathrm{d}t}{\Pr(X > u|\mathbf{Z}_i)},$$

respectively. Furthermore, in the partial likelihood, all that matters is the $\mathbf{Z}$ values of the members of the risk set at each failure time, not whether a given individual happens to appear in several different risk sets. Thus, if $\mathbf{Z}(t)$ is only updated at a few times per person, it is simplest to treat each person as several "individuals" each with a time fixed covariate. Let the vector $(T_0, T, D, \mathbf{Z})$ denote the entry and exit times, the censoring indicator, and the value of $\mathbf{Z}(t)$ for $t \in (T_0, T]$, respectively. Then an individual who enters a trial at time 0 with $Z(t) = -2$ for $0 \le t \le 1$, $Z(t) = -3$ for $1 < t \le 2$, $Z(t) = 2.5$ for $2 < t \le 3$, and $Z(t) = 2$ for $3 < t \le 3.6$ and dies at $T = 3.6$ is represented by the four data points (0, 1, 0, −2), (1, 2, 0, −3), (2, 3, 0, 2.5), and (3, 3.6, 1, 2).

When computing the likelihood with fixed regressors, it makes sense to use an updating formula. As one moves from one time point to the next, the risk set changes slightly due to the entry or the exit (due to death or censoring) of "individuals". The values for those "individuals" who remain in the risk set do not change and need not be recalculated. In this way the calculation is kept to order n (albeit $4n$ if each individual is treated as four because of changing covariate values).

By contrast, when using continuously varying regressors, one has no choice but to recalculate the partial likelihood contribution from each time point from scratch. This makes the calculation order n^2.

Many software packages that will handle updated regressors will not (easily) handle continuously varying regressors. It is difficult to fit models with user-defined variables such as $Z_2(t) = tZ_1$ using such packages. One might wish to compare the models with hazards $\lambda_0(t)\exp(\beta_1 Z_1)$ and $\lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 t Z_1)$. Of course, for the purpose of testing $\beta_2 = 0$, it is not necessary to fit the latter model. Instead, one may calculate the score statistic for $\beta_2 = 0$ evaluated at the maximum partial likelihood estimate of β_1 from the model with the single (fixed) regressor.

Model Checking

An important aspect of modeling any set of data is assessing the adequacy of the fit and checking to see that the resulting inference is not unduly influenced by a few observations. In general the iterative process of model building and checking may be considered an art rather than a science. Here we review some of the tools available to the statistical artisan analyzing survival data by means of a Cox model.

The simplest form of graphical check comes from dividing the data into groups based on some explanatory variable and fitting a stratified Cox model. If the explanatory variable "$Z = s$" is well modeled by the Cox model, one has $\Lambda_{s0} = \Lambda_0 \exp(\gamma s)$, say. Thus, plotting the logarithm of the cumulative hazard estimate from each strata should reveal parallel curves. That is, the vertical distance between the two curves $\log \Lambda_{r0}(t)$ and $\log \Lambda_{s0}(t)$ should be the same for all t. The common distance should be $\gamma(r - s)$. In practice, such graphics, while intuitively appealing, are not particularly useful.

A closely related, but rather more useful, graph for two strata is obtained by plotting one cumulative hazard $\Lambda_{r0}(t)$ against the other $\Lambda_{s0}(t)$ for all or some selected values of t. Under proportional hazards, such an H–H plot should approximate a straight line through the origin with slope $\exp(\gamma s)$ [6, Section VII.3.1]. The method is easily extended to multiple strata. The disadvantages of the H–H plot are that they do not record the actual time t, and that, if the proportional hazards assumption is seen to be violated, it is difficult to know how to modify the proportional hazards model other than by using a stratified model. Hess [42] reviews a number of variants on these two simple graphical checks of proportional hazards and compares eight graphical methods on each of three data sets. He recommends smoothed plots of scaled Schoenfeld residuals. These are described in the subsection on residuals.

Goodness-of-Fit Tests

Several authors have developed formal goodness-of-fit tests. These can be divided into those designed to be able to detect global alternatives and those with greater power at detecting some specified alternative. Virtually all the tests are asymptotically equivalent to tests based on a defined time-dependent explanatory variable. We saw earlier that the first such tests were proposed by Cox himself [21]. One may add an additional regressor $Z_*(t) = Zg(t)$ for some function $g(t)$. Common choices for g included the identity function $g(t) = t$ and its logarithmic transform $g(t) = \log t$. Other authors use step functions that may jump at either a fixed or a random (but predictable) time. If the partial likelihood is maximized with $Z(t)$, then the partial likelihood ratio test is the statistic of choice. But for testing the goodness of fit, the score test is simpler to compute because it does not require fitting a model with a time-dependent regressor.

Gill & Schumacher [34] proposed a family of tests of the proportional hazards assumption between two samples, A and B. Their tests are motivated by comparing two different estimates of the relative hazard between the two samples. Under proportional hazards the two estimates will be similar, but they need not be in general. The estimates of relative hazard used are derived from **linear rank tests**, which are themselves equivalent to score tests from the Cox partial likelihood with specially defined time-dependent regressors [33]. The family of tests proposed by Gill & Schumacher [34] are thus similar in spirit to those proposed by Breslow et al. [14]. The latter consider the score test for $\beta_2 = 0$ in the model

$$\lambda_B(t) = \lambda_A(t) \exp[\beta_1 + \beta_2 g(t)],$$

corresponding to covariates $Z_1 = I(B)$ and $Z_2(t) = I(B)g(t)$. A popular choice is $g(t) = \hat{S}(t)$, the Kaplan–Meier estimate of survival in the combined sample at t. O'Quigley & Pessione [62] suggest using a step function for $g(t)$. For a one degree of freedom test, one must choose both the cut points and the values of the step function. For a more general alternative hypothesis, one could partition the time axis into J intervals. The null hypothesis is that the relative hazards $\exp \beta_j$ in all $j = 1, \ldots, J$ intervals are the same, and this can be tested with $J - 1$ degrees of freedom. Wei [85] proposes an omnibus goodness-of-fit test for the two-sample problem based on the supremum of the score statistic $\sup_t |U(\hat{\beta}, t)|$.

Schoenfeld [75] was interested in a more general goodness-of-fit test for the Cox model. He suggested embedding a Cox model with regressor Z in a much larger model with regressors Z and $\mathbf{Z}_*(t)$, where the $\mathbf{Z}_*(t)$ are a set of indicator variables that partition the regressor–time space. Thus, for instance, one might divide the time axis into three parts and the covariate space into four, and form the Cartesian product with 12 cells. In addition to the score test for the coefficients of $\mathbf{Z}_*$ being all zero, one can examine the "residuals", i.e. the difference between the observed and expected (under the basic model with covariate Z) number of events in each of the 12 cells. Lin et al. [58] avoid the need for an arbitrary partition of the space by deriving a supremum test based on the cumulative sum

$$W(t, z) = \sum_{Z_i \le z} [O_i(t) - E_i(t)],$$

where $O_i(t) = N_i(t)$ and $E_i(t) = \int_0^t Y_i(u)\, \mathrm{d}\hat{\Lambda}_i(t)$ are, respectively, the observed and expected number of events in individual i, by time t.

Residuals

There have been numerous attempts to define residuals and to propose diagnostic plots for the Cox model

(*see* **Diagnostics**). The situation is complicated by both the **semiparametric model** and the presence of censoring. Some of the proposed techniques are decidedly less useful than one might have hoped. In particular, attempts to define residuals that (under the Cox model) look like a random sample from a specified distribution, so that $Q-Q$ plots can be drawn (*see* **Normal Scores**), have failed. Graphical assessment of the functional form of a covariate and of the constancy of the regression parameters over time have been more successful.

An early definition of residual for the Cox model was the estimated cumulative intensity for each individual:

$$\hat{A}_i(\infty) = \int_0^\infty Y_i(u)\,\mathrm{d}\hat{\Lambda}_i(u) = \int_0^\infty Y_i(u)\exp[\hat{\boldsymbol{\beta}}'\mathbf{z}_i(u)]\,\mathrm{d}\hat{\Lambda}_0(u) \quad (11)$$

[if $Y_i(u) = I(T_i \geq u)$ and $\mathbf{z}_i(u) = \mathbf{z}_i$, then $\hat{A}_i(\infty) = \hat{\Lambda}_i(T_i) = \exp(\hat{\boldsymbol{\beta}}'\mathbf{z}_i)\hat{\Lambda}_0(T_i)$] [24, 52]. Later authors made an adjustment to the residual depending on whether the individual was censored or not. The resulting residual $r_i = D_i - \hat{A}_i(\infty)$ is called the martingale residual and is a special case of the general family of residual processes defined by

$$\int_0^t H_i(u)\,\mathrm{d}\hat{M}_i(t), \quad (12)$$

where $\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(u)\,\mathrm{d}\hat{\Lambda}_i(u)$ and H_i is a predictable process [8, 81]. Thus r_i is the estimated martingale transform, (12), with $H_i = 1$ and $t = \infty$. The martingale residual may be thought of as the difference between the observed and the expected number of events for the ith individual. The distribution of martingale residuals in a survival setting is very skewed since they have mean zero (under the true model) but range from 1 (for someone who fails at time 0) to minus a very large number (for someone who survives much longer than "expected"). Summing over individuals with similar covariate values $\{i : \mathbf{z}_i \in \mathcal{Z}\}$, say, one obtains the residual number of events for individuals with $\mathbf{z} \in \mathcal{Z}$. Thus, smoothing the martingale residuals against a regressor (or a potential regressor) gives an indication as to how well the model fits the data. Systematic departures from zero indicate that there is an excess (or deficit) in the modeled hazard for that group of individuals. Heuristically one has

$$\mathbf{E}\{N_i(\infty)|\mathbf{z},z^*\} = A(\infty|\mathbf{z}, z^*) \approx \hat{A}(\infty|\mathbf{z}) + \text{smooth}(r_i|z^*).$$

More recently, Grambsch et al. [36] have considered the model

$$\lambda(t|\mathbf{z},z^*) = \lambda_0(t)\exp[\boldsymbol{\beta}'\mathbf{z} + f(z^*)]. \quad (13)$$

They propose fitting the Cox model with prognostic index $\boldsymbol{\beta}'\mathbf{z} + \gamma z^*$ and plotting $\log\{\text{smooth}[N_i(\infty)]\} - \log\{\text{smooth}\,[\hat{A}_i(\infty)]\} + \hat{\gamma}z^*$ vs. z^*. The smooth curve will approximate $f(z^*)$ to first order. In practice, the approximation seems to work well even when Z^* is correlated with the other regressors $\mathbf{Z}$.

The martingale residuals were defined by integrating the martingale difference array $\mathrm{d}\hat{M}_i(t)$ over the time axis to give a single residual per individual. To examine the proportional hazards assumption, one is more interested in obtaining a separate residual for each failure time. This can be done by summing the martingale differences, at a given time, over all individuals. Now $\Sigma_i\,\mathrm{d}\hat{M}_i(t) = 0$ for all t by the definition of the Breslow estimator $\hat{\Lambda}_0$. Nevertheless, one can use the martingale transform, (12), with $\mathbf{H}_i = \mathbf{Z}_i$. Then at each failure time one is comparing the observed value of $\mathbf{Z}$ in the individual that fails with its expected value. Such a residual,

$$\mathbf{r}^*(T_j) = \sum_i \mathbf{Z}_i(T_j)[\,\mathrm{d}N_i(T_j) - Y_i(T_j)\,\mathrm{d}\hat{\Lambda}_i(T_j)] = \mathbf{S}_j - d_j\frac{\mathbf{S}^{(1)}(\hat{\boldsymbol{\beta}}, T_j)}{\mathbf{S}^{(0)}(\hat{\boldsymbol{\beta}}, T_j)},$$

was first proposed by Schoenfeld [76]. It is seen that the sum of the Schoenfeld residuals evaluated at $\boldsymbol{\beta}$ is equal to the score $\mathbf{U}(\boldsymbol{\beta})$. It is not difficult to show that, even under the model

$$\lambda(t|\mathbf{z}) = \lambda_0(t)\exp[\boldsymbol{\beta}(t)'\mathbf{z}], \quad (14)$$

$\mathbf{S}^{(1)}\,[\boldsymbol{\beta}(t), t]/\boldsymbol{S}^{(0)}\,[\boldsymbol{\beta}(t), t] \rightarrow \mathbf{E}(\mathbf{Z}|T = t,\ D = 1)$. Thus, using a one-step Taylor series expansion about $\boldsymbol{\beta}(t) = \hat{\boldsymbol{\beta}}$, one has

$$\boldsymbol{\beta}(t) \approx \hat{\boldsymbol{\beta}} + \hat{\mathbf{V}}(t)^{-1}\mathbf{r}^*(t),$$

where

$$\hat{\mathbf{V}}(t) = \frac{\mathbf{S}^{(2)}(\hat{\boldsymbol{\beta}}, t)}{\mathbf{S}^{(0)}(\hat{\boldsymbol{\beta}}, t)} - \left(\frac{\mathbf{S}^{(1)}(\hat{\boldsymbol{\beta}}, t)}{\mathbf{S}^{(0)}(\hat{\boldsymbol{\beta}}, t)}\right)^{\otimes 2}.$$

Hence, Grambsch & Therneau [35] have proposed plotting a smooth of $\hat{\boldsymbol{\beta}} + \hat{\mathbf{V}}(t)^{-1}\mathbf{r}^*(t)$ against t in order to get a feel of $\boldsymbol{\beta}(t)$. Often $\mathbf{V}(t) = \lim_{n\to\infty} \hat{\mathbf{V}}(t)$ does not vary much as a function of t, so for exploratory purposes it may be enough to use $\mathbf{I}(\hat{\beta})/\Sigma N_i(\infty)$ in place of $\hat{\mathbf{V}}(t)$. This has the advantage of not having to store and invert a different covariance matrix at each failure time. In practice, $\mathbf{V}(t)$ will vary most when a variable Z has a skewed distribution and those in the tail are at greatest risk. In all cases it will be difficult to estimate $\hat{\mathbf{V}}(t)$ if the risk set is small at time t, and it is also for large values of t that the $V(t)$ is most likely to be substantially different from its average value.

Influence Diagnostics

Various measures of influential observations have been suggested for the Cox model. The influence **diagnostic** is intended to approximate the amount by which the regression estimate $\hat{\boldsymbol{\beta}}$ would change if the ith individual were removed from the data set [69]. One such approximation is the infinitesimal **jackknife**, first proposed by Cain & Lange [17]. Their residuals are equal to the components of the scaled efficient score statistic. This can be written as a martingale transform residual with $\mathbf{H}_i(t) = \mathbf{Z}_i - \mathbf{E}(\hat{\boldsymbol{\beta}}, t)$. The scaling is done by $\mathbf{I}(\hat{\boldsymbol{\beta}})^{-1}$. One has

$$\tilde{\mathbf{r}}_i = I(\hat{\boldsymbol{\beta}})^{-1} \int_0^\infty [\mathbf{z}_i - \mathbf{E}(\hat{\boldsymbol{\beta}}, t)]\, \mathrm{d}\hat{M}_i(t).$$

An alternative estimate of the influence of an individual is given by Storer & Crowley [79].

Alternatives and Extensions

We have already discussed many extensions of the basic Cox model. We have permitted nonproportional hazards through the stratified Cox model and through user-defined time-dependent variables. We have considered diagnostics to detect data that appear to come from more nonparametric models, such as the additive Cox model $\lambda(t|\mathbf{z}) = \lambda_0(t)\exp[\sum_k f_k(z_k)]$, in which some of the functions f_k may be assumed to be linear while others are left unspecified [32, 37, 39, 70], and the multiplicative hazards model $\lambda(t|\mathbf{z}) = \lambda_0(t)\exp[\boldsymbol{\beta}(t)'\mathbf{z}]$ [30, 40, 41, 86, 84].

We have also seen how the model that was originally perceived for survival data can be generalized quite naturally to event data in which a single individual may have multiple events. The events need not even all be of the same type. They may represent competing risks or more generally the various states in a multistate model. In the classic heart transplant situation, for example, one might use Cox regression to model the transition from identification as a potential recipient (state 0) to transplant (state 1); from state 0 to death (state 2); and from state 1 to death [26]. Three state models in which transitions from state 1 (diseased) back to state 0 (healthy) are possible are also common (see, for example, Andersen et al. [6, Example VII.2.10]). The study of (i) acute graft-vs.-host disease, (ii) chronic graft-vs.-host disease, (iii) leukemia relapse, and (iv) death following bone marrow transplantation (state 0) is also considered by Andersen et al. [6, Example VII.2.18].

We briefly mention a few alternatives to the semiparametric Cox model for regression analysis of censored survival data. Naturally one can try to adapt estimation in any parametric regression model to cope with right censored data. Loglinear models with **Weibull** or **Gamma** errors [50, Section 3.6] tend to be more popular in reliability (engineering) than in biostatistics. Particularly in epidemiology, one sometimes has a known population mortality rate that one wants to use in place of the baseline hazard function. The hazard for individual i is given by

$$\lambda_i(t) = \mu_i(t)\exp[\boldsymbol{\beta}'\mathbf{Z}_i(t)],$$

where μ_i is the population mortality corresponding to individual i [7, 15]. Fully parametric models have been studied using counting process techniques by Borgan [12]. Another Cox-like model that uses a known rate is the proportional excess hazards model [73] (*see* **Excess Risk**),

$$\lambda_i(t) = \mu_i(t) + \lambda_0(t)\exp[\boldsymbol{\beta}'\mathbf{Z}_i(t)],$$

in which the excess mortality is modeled by a Cox model.

A general family, known as the **accelerated failure-time model**, is a linear regression model for

the logarithm of the survival time,

$$\log X = \boldsymbol{\beta}'\mathbf{Z} + \varepsilon,$$

where the error ε may be either from a specified distribution or from an unknown distribution. Gaussian errors and no censoring simply correspond to linear regression of $\log X$. If the errors are Weibull, then the model is also a proportional hazards model. Theoretical attention has focused on the semiparametric model with unknown error distribution. Another family of models that include the Cox model as a special case are the transformation models in which an unknown monotone transformation of the survival time is assumed to have a linear regression:

$$\psi(X) = \boldsymbol{\beta}'\mathbf{Z} + \varepsilon.$$

If the error distribution is **extreme value** ($\exp\varepsilon$ distributed exponential with mean 1), then the transformation model is a Cox model with cumulative baseline hazard given by $\exp[\psi(t)]$ and regression parameter $s - \boldsymbol{\beta}$.

The Cox model is a multiplicative hazards model. Aalen [2] introduced an additive hazards model (*see* **Aalen's Additive Regression Model**). A semiparametric version of the model [61] is given by

$$\lambda(t|\mathbf{Z}_1, \mathbf{Z}_2) = \boldsymbol{\theta}_1(t)'\mathbf{Z}_1(t) + \boldsymbol{\theta}_2'\mathbf{Z}_2(t).$$

If the variables $\mathbf{Z}_1$ include a constant, then we may pull out a baseline hazard and write the first term on the right of the equation as $\lambda_0(t) + \alpha_{11}(t)Z_{11}(t) + \cdots + \alpha_{1p}(t)Z_{1p}(t)$. The cumulative components of hazard $A_{1j}(t) = \int_0^t \alpha_{1j}(u)\,\mathrm{d}u$ can be estimated at parametric rates, and these must be smoothed to estimate the $\alpha_{1j}(u)$. The model extends naturally to the more general counting process formulation.

Several authors have considered "special" Cox models. These include models for matched pairs [38, 44] (*see* **Matching**), and for **interval censored** survival data [46], a model for periodic data [64] (*see* **Seasonal Time Series**), and a model for **case–cohort** data [66, 77]. **Bayesian analysis** of the Cox model was first considered by Kalbfleisch [49; [50], Section 8.4] and later by Hjort [43].

There has been relatively little written about robust estimation in the Cox model (*see* **Robustness**). Estimators that maximize a weighted partial likelihood have been proposed independently at least three times [56, 71, 72, 74]. The weights may be random and may depend on the regressors $\mathbf{Z}$, but they should be predictable (or at least asymptotically equivalent to predictable weights). A slightly different estimator which essentially corresponds to the efficient score function from a weighted full likelihood has also been studied [10].

Consideration of the Cox estimator for $\hat{\boldsymbol{\beta}}$ when the data do not come from a Cox model leads naturally to adoption of the sandwich estimator of the variance of $\hat{\boldsymbol{\beta}}$ [57, 69]. This is the usual infinitesimal jackknife estimator that can be obtained from the influence residuals

$$\tilde{\mathrm{var}}(\hat{\boldsymbol{\beta}}) = \sum_i \tilde{r}_i \tilde{r}_i'.$$

The estimator is perhaps most useful when the data are clustered (*see* **Clustering**). Suppose that $\tilde{\mathbf{r}}_{ki}$ is the influence residual from individual i in cluster k. Then define $\tilde{\mathbf{r}}_k = \Sigma_i \tilde{\mathbf{r}}_{ki}$ and estimate the variance of $\hat{\boldsymbol{\beta}}$ by $\Sigma_k \tilde{\mathbf{r}}_k^{\otimes 2}$ [55]. This may be a simple technique for adjusting inference when using the Cox model with multivariate survival data. For instance, if each person could have several events, then one might wish to treat the person as a cluster. In another example, the clusters might be formed from survival data on individuals within families.

Another approach adapting the Cox model to multivariate data is through latent variables or **frailties**. The idea is that, conditionally on an unobserved variable or frailty, the survival times follow a Cox model. The value of the frailty W_i is assumed to be the same for all survival times within a cluster. Two frailty distributions have received the most attention: Clayton & Cuzick [19] considered the hazard model $\lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{Z} + W)$ in which $\exp W$ has a gamma distribution; Hougaard [45] favors using the positive stable distribution, as this is the only choice that yields proportional hazards both marginally (integrating over the unobserved variable) and conditionally.

References

[1] Aalen, O.O. (1978). Nonparametric inference for a family of counting processes, *Annals of Statistics* **6**, 701–726.

[2] Aalen, O.O. (1980). A Model for Nonparametric Regression Analysis of Counting Processes, *Springer Lecture Notes in Statistics*, Vol. 2. Springer-Verlag, New York, pp. 1–25.

[3] Altman, D.G. & De Stavola, B.L. (1994). Practical problems in fitting a proportional hazards model to data

with updated measurements of the covariates, *Statistics in Medicine* **13**, 301–341.

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

[5] Andersen, P.K. & Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study, *Annals of Statistics* **10**, 1100–1120.

[6] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[7] Andersen, P.K., Borch-Johnsen, K., Deckert, T., Green, A., Hougaard, P., Keiding, N. & Kreiner, S. (1985). A Cox regression model for the relative mortality and its application to diabetes mellitus survival data, *Biometrics* **41**, 921–932.

[8] Barlow, W.E. & Prentice, R. (1988). Residuals for relative risk regression, *Biometrika* **75**, 65–74.

[9] Bednarski, T. (1989). On sensitivity of Cox's estimator, *Statistics and Decisions* **7**, 215–228.

[10] Bednarski, T. (1993). Robust estimation in Cox's regression model, *Scandinavian Journal of Statistics* **20**, 213–225.

[11] Begun, J.M., Hall, W.J., Huang, W.M. & Wellner, J.A. (1983). Information and asymptotic efficiency in parametric - nonparametric models, *Annals of Statistics* **11**, 432–452.

[12] Borgan, Ø. (1984). Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data, *Scandinavian Journal of Statistics* **11**, 1–16. Correction **11** (1984) 275.

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

[14] Breslow, N.E., Edler, L. & Berger, J. (1984). A two-sample censored-data rank test for acceleration, *Biometrics* **40**, 1049–1062.

[15] Breslow, N.E., Lubin, J.H., Marek, P. & Langholz, B. (1983). Multiplicative models and cohort analysis, *Journal of the American Statistical Association* **78**, 1–12.

[16] Bretagnolle, J. & Huber-Carol, C. (1988). Effects of omitting covariates in Cox's regression model for survival data, *Scandinavian Journal of Statistics* **15**, 125–138.

[17] Cain, K.C. & Lange, N.T. (1984). Approximate case influence for the proportional hazards regression model with censored data, *Biometrics* **40**, 493–499.

[18] Christensen, E., Neuberger, J., Crowe, J., Altman, D.G., Popper, H., Portmann, B., Doniach, D., Ranek, L., Tygstrup, N. & Williams, R. (1985). Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis: final results of an international trial, *Gastroenterology* **89**, 1084–1091.

[19] Clayton, D. & Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion), *Journal of the Royal Statistical Society, Series A* **148**, 82–117.

[20] Collett, D. (1994). *Modelling Survival Data in Medical Research*. Chapman & Hall, London.

[21] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[22] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[23] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[24] Cox, D.R. & Snell, E.J. (1968). A general definition of residuals (with discussion), *Journal of the Royal Statistical Society, Series B* **30**, 248–275.

[25] Crowley, J. & Hu, M. (1977). Covariance analysis of heart transplant survival data, *Journal of the American Statistical Association* **72**, 27–36.

[26] Crowley, J. & Storer, B.E. (1983). Comment on "A reanalysis of the Stanford heart transplant data", by Aitkin, Laird and Francis, *Journal of the American Statistical Association* **78**, 277–281.

[27] Efron, B. (1977). The efficiency of Cox's likelihood function for censored data, *Journal of the American Statistical Association* **72**, 557–565.

[28] Feigl, P. & Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information, *Biometrics* **21**, 826–838.

[29] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[30] Gamerman, D. (1991). Dynamic Bayesian models for survival data, *Applied Statistics* **40**, 63–79.

[31] Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples, *Biometrika* **52**, 203–223.

[32] Gentleman, R. & Crowley, J. (1991). Local full likelihood estimation for the proportional hazards model, *Biometrics* **47**, 1283–1296.

[33] Gill, R.D. (1984). Understanding Cox's Regression Model: a martingale approach, *Journal of the American Statistical Association* **79**, 441–447.

[34] Gill, R.D. & Schumacher, M. (1987). A simple test for the proportional hazards assumption, *Biometrika* **74**, 289–300.

[35] Grambsch, P.M. & Therneau, T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals, *Biometrika* **81**, 515–526.

[36] Grambsch, P.M., Therneau, T.M. & Fleming, T.R. (1995). Diagnostic plots to reveal functional form for covariates in multiplicative intensity models, *Biometrics* **51**, 1469–1482.

[37] Gray, R.J. (1992). Flexible methods for analysing survival data using splines, with applications to breast cancer prognosis, *Journal of the American Statistical Association* **87**, 942–951.

[38] Gross, S.T. & Huber, C. (1987). Matched pair experiments: Cox and maximum likelihood estimation, *Scandinavian Journal of Statistics* **14**, 27–41.

[39] Hastie, T. & Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model, *Biometrics* **46**, 1005–1016.

[40] Hastie, T. & Tibshirani, R. (1993). Varying coefficient models (with discussion), *Journal of the Royal Statistical Society, Series B* **55**, 757–797.

[41] Hess, K.R. (1994). Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions, *Statistics in Medicine* **13**, 1045–1062.

[42] Hess, K.R. (1995). Graphical methods for assessing violations of the proportional hazards assumption in Cox regression, *Statistics in Medicine* **14**, 1707–1723.

[43] Hjort, N.L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data, *Annals of Statistics* **18**, 1259–1294.

[44] Holt, J.D. & Prentice, R.L. (1974). Survival analysis in twin studies and matched pair experiments, *Biometrika* **65**, 159–166.

[45] Hougaard, P. (1984). Life table methods for heterogeneous populations: distributions describing the heterogeneity, *Biometrika* **71**, 75–83.

[46] Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring, *Annals of Statistics* **24**, 540–568.

[47] Jacobsen, M. (1989). Existence and unicity of MLEs in discrete exponential family distributions, *Scandinavian Journal of Statistics* **16**, 335–349.

[48] Johansen, S. (1983). An extension of Cox's regression model, *International Statistical Review* **51**, 165–174.

[49] Kalbfleisch, J.D. (1978). Nonparametric Bayes analysis of survival data, *Journal of the Royal Statistical Society, Series B* **40**, 214–221.

[50] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[51] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[52] Kay, R. (1977). Proportional hazards regression models and the analysis of censored survival data, *Applied Statistics* **26**, 227–237.

[53] Kleinbaum, D.G. (1995). *Survival Analysis: A Self-Learning Text*. Springer-Verlag, New York.

[54] Lagakos, S. & Schoenfeld, D. (1984). Properties of proportional hazards score tests under misspecified regression models, *Biometrics* **40**, 1037–1048.

[55] Lee, E.W., Wei, L.J. & Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations, in *Survival Analysis: State of the Art*, J.P. Klein, & P.K. Goel, eds. Kluwer, Dordrecht, pp. 237–247.

[56] Lin, D.Y. (1991). Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators, *Journal of the American Statistical Association* **86**, 725–728.

[57] Lin, D.Y. & Wei, L.J. (1989). The robust inference for the Cox proportional hazards model, *Journal of the American Statistical Association* **84**, 1074–1078.

[58] Lin, D.Y., Wei, L.J. & Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residual, *Biometrika* **80**, 557–572.

[59] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer and Chemotherapy Reports* **50**, 163–170.

[60] Marubini, E. & Valsecchi, M.G. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies*. Wiley, Chichester.

[61] McKeague, I. & Sasieni, P. (1994). A partly parametric additive risk model, *Biometrika* **81**, 501–514.

[62] O'Quigley J. & Pessione F. (1989). Score tests for homogeneity of regression effects in the proportional hazards model, *Biometrics* **45**, 135–144.

[63] Peto, R. (1972). Contribution to the discussion of paper by D.R. Cox, *Journal of the Royal Statistical Society, Series B* **34**, 205–207.

[64] Pons, O. & de Turckheim, E. (1988). Cox's periodic regression model, *Annals of Statistics* **16**, 678–693.

[65] Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model, *Biometrika* **69**, 331–342.

[66] Prentice, R.L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials, *Biometrika* **73**, 1–11.

[67] Prentice, R. & Self, S. (1983). Asymptotic distribution theory for Cox-type regression models with general risk form, *Annals of Statistics* **11**, 804–813.

[68] Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions, *Annals of Statistics* **11**, 453–466.

[69] Reid, N. & Crépeau, H. (1985). Influence functions for proportional hazards regression, *Biometrika* **72**, 1–9.

[70] Sasieni, P. (1992). Information bounds for the conditional hazard ratio in a nested family of regression models, *Journal of the Royal Statistical Society, Series B* **54**, 617–635.

[71] Sasieni, P.D. (1993). Maximum weighted partial likelihood estimates for the Cox model, *Journal of the American Statistical Association* **88**, 144–152.

[72] Sasieni, P.D. (1993). Some new estimates for Cox regression, *Annals of Statistics* **21**, 1721–1759.

[73] Sasieni, P.D. (1996). Proportional excess hazards, *Biometrika* **83**, 127–141.

[74] Schemper, M. (1992). Cox analysis of survival data with non proportional hazards functions, *Statistician* **41**, 455–465.

[75] Schoenfeld, D. (1980). Chi-squared goodness of fit tests for the proportional hazards regression model, *Biometrika* **67**, 145–153.

[76] Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model, *Biometrika* **69**, 239–241.

[77] Self, S.G. & Prentice, R.L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies, *Annals of Statistics* **16**, 64–81.

[78] Solomon, P.J. (1984). Effect of misspecification of regression models in the analysis of survival data, *Biometrika* **71**, 291–298. Amendment. **73** (1986) 245.

[79] Storer, B.E. & Crowley, J. (1985). A diagnostic for Cox regression and general conditional likelihoods, *Journal of the American Statistical Association* **80**, 139–147.

[80] Struthers, C.A. & Kalbfleisch, J.D. (1986). Misspecified proportional hazards models, *Biometrika* **73**, 363–369.

[81] Therneau, T.M., Grambsch, P.M. & Fleming, T.R. (1990). Martingale-based residuals for survival models, *Biometrika* **77**, 147–160.

[82] Thomas, D.C. (1981). General relative risk models for survival time and matched case-control analysis, *Biometrics* **37**, 673–686.

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

[84] Verweij, P.J.M. & van Houwelingen, H.C. (1995). Time-dependent effects of fixed covariates in Cox regression, *Biometrics* **51**, 1550–1556.

[85] Wei, L.J. (1984). Testing goodness of fit for the proportional hazards model with censored observations, *Journal of the American Statistical Association* **80**, 139–147.

[86] Zucker, D.M. & Karr, A.F. (1990). Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach, *Annals of Statistics* **18**, 329–353.

PETER SASIENI