Tests for the Proportional Hazards Assumption in the Cox Model

Linda Reiersølmoen
Department of Mathematical Sciences
Norwegian University of Science and Technology
N-7491 Trondheim, Norway
E-mail:lindare@math.ntnu.no

Jan Terje Kvaløy
Department of Mathematics and Science
Stavanger University College
N-4091 Stavanger, Norway
E-mail:jan.t.kvaloy@tn.his.no

January 12, 2001

Abstract

New test statistics to test the proportional hazards assumption in the Cox model are proposed. The tests are based on the score process. If the score process cannot be accepted as a Brownian bridge, the null hypothesis of proportional hazards has to be rejected. We consider a previously proposed Kolmogorov-Smirnov type statistic in addition to the proposed tests based on Cramér-von Mises and Anderson-Darling type statistics. These new tests are supposed to have power to detect both monotone and non-monotone alternatives. For comparison we also consider a chi-square test based on the Schoenfeld residuals.

Simulation studies indicate that the proposed tests have a more correct level than the Kolmogorov-Smirnov test and the test based on the Schoenfeld residuals. For monotone alternatives they also have a higher power than the previously proposed tests. For the non-monotone alternative studied here, however, it seems like the Kolmogorov-Smirnov test has a somewhat higher power. In this case the chi-square test has almost no power. It is also recommended to plot the score process. This can give useful information about tendencies in the data. The application of the tests on real data are also demonstrated.

KEY WORDS: Cox model; Tests for proportional hazards; Level properties; Power properties; Censored data

1 Introduction

In the study of time event data, Cox's proportional hazards model (Cox, 1972) is often used. An event is a qualitative change that can be situated in time, e.g. death or failure of a technical component. Often, only events that can happen once (e.g. death) are considered. Then the time from a subject starts to be under risk of the event considered in the study until it e.g. dies or fails, T, is called time to failure or also survival time. Sometimes, however, recurrent events may occur and then T represents the event times. Both single and recurrent event situations can be modeled as a counting process, $\{N(t), t \geq 0\}$. Much of the background material used in this paper is taken from Andersen, Borgan, Gill and Keiding (1993).

A feature often present in survival data is known as censoring. Broadly speaking, censoring occurs when some event times are only known to have occurred within certain intervals, while the remainder of the event times are known exactly. Right censoring is the most common form for such 'incomplete observation' and the only form we will concentrate on here. Right censoring means that the observation of units is ceased after some (possibly random) time, i.e. the events are observed only if they occur prior to some censoring time. In recurrent event cases, the right censoring expresses how the observation of each process is terminated. In our further discussion, independent and non-informative censoring is assumed (Andersen et al., 1993).

For the case with only one event, the observation time is the minimum of the survival time and the censoring time. In this case the observation times will be indexed by a simple subscript i, $T_i,\ i = 1, \ldots, n$, where n is the number of observations. For recurrent events the observation times are the event times until the process is censored, and the last observation time for the process is the censoring time of the process. For these processes the event times will be indexed by a double subscript ij, $T_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, N_i(\tau_i)$, where n is the number of processes and $N_i(\tau_i)$ is the number of observed events in process i (τ_i is the censoring time for process i, i.e. the time when the observation of process i is terminated). See Figure 1 for an example.

The intensity process of a counting process is under reasonable assumptions defined as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(N(t + \Delta t) - N(t) = 1 | \mathcal{F}_{t^-})}{\Delta t}, \tag{1}$$

where $\mathcal{F}_{t^-}$ is the history of the process until (not included) time t, i.e. $\mathcal{F}_{t^-}$ represents the available data just before time t. The cumulative intensity is given by

$$\Lambda(t) = \int_0^t \lambda(u) du \tag{2}$$

A counting process N behaves locally at time t, conditionally on the past, just like a Poisson process with rate $\lambda(t)$ (Andersen et al., 1993). In the present paper it is further

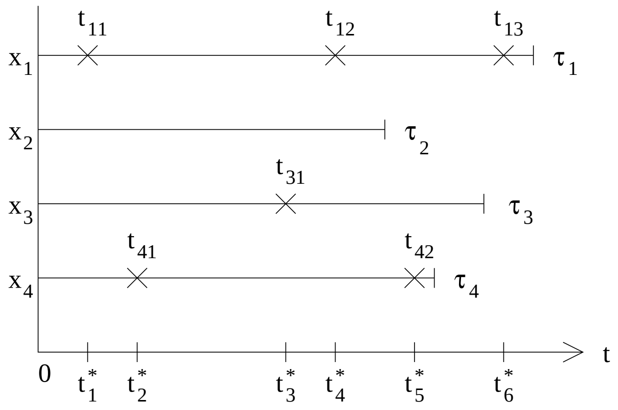

Figure 1: Example of time event data for counting processes.

assumed that the counting processes under consideration are non-homogeneous Poisson processes (NHPP) and hence that the conditional intensity, $\lambda(t)$, is deterministic.

Cox's regression model describes for this case the intensity $\lambda(t)$ as a function of a set of covariates which is expected to influence this intensity. More precisely the Cox model is

$$\lambda(t|\mathbf{x}) = \lambda_0(t)e^{\boldsymbol{\beta}^T\mathbf{x}}, \tag{3}$$

where $\lambda(t|\mathbf{x})$ is the intensity of the process, $\mathbf{x}$ is the vector of covariates for the process, and $\boldsymbol{\beta}$ is a vector of regression parameters. In the general case the covariate vector, $\mathbf{x}$, can depend on time, but we shall mainly consider the situation when $\mathbf{x}$ is constant over time. Thus, the model assumes a log-linear relation between the intensity and the covariates. The baseline intensity, $\lambda_0(t)$, is the intensity if all the covariates have the value zero and is in the Cox model a completely unspecified function. It is, however, assumed to be non-negative with $\int_0^t \lambda_0(u)du < \infty$. This baseline intensity, $\lambda_0(t)$, is assumed to be common for all processes. From a partial likelihood function, an estimator for $\boldsymbol{\beta}$ can be defined independently of $\lambda_0(t)$ (Cox, 1975). The baseline intensity can then be estimated by a non-parametric method. The model is therefore called semi-parametric.

If there is reason to believe that the observed processes are governed by k different distributions but that the covariates have the same multiplicative influence on the intensities, a stratified Cox model can be used for each process, in the jth stratum:

$$\lambda_j(t|\mathbf{x}) = \lambda_{j0}(t)e^{\boldsymbol{\beta}^T\mathbf{x}},\ j = 1, \ldots, k \tag{4}$$

There exist several types of residuals, plots and tests to investigate model fit and to check the assumptions of the Cox model. In this paper we will especially look at the assumption of proportional hazards, or proportional intensities, i.e. the assumption that the intensities' dependency on the covariates is multiplicative and independent of time.

One possible method to check the proportional hazards assumption is to use the Schoenfeld residuals (Schoenfeld, 1982). If the influence on the intensity, of covariate x_i, changes over time, a possible model could be

$$\lambda(t|\mathbf{x}) = \lambda_0(t)e^{\boldsymbol{\beta}^T\mathbf{x}+\theta g(t)x_i}, \tag{5}$$

where $g(t)$ is varying about zero. A smoothed plot of the Schoenfeld residuals with respect to this covariate against time, should give a picture of the function $g(t)$. (See Figure 4 for an example.) A chi-square test corresponding to these residuals to test if $\theta = 0$ or not was derived by Grambsch and Therneau (1994). This method may, however, not be appropriate to test for non-monotonic time dependencies, i.e. non-monotonic $g(t)$.

There is also a purely graphical method to check whether $\theta = 0$. The data are divided into several strata, see (4), and models are fitted, with different $\lambda_{j0}(t)$ but equal $\boldsymbol{\beta}$ for all strata. Under the proportional hazards assumption, a plot of $\log(\hat{\Lambda}_{10}(t)), \ldots, \log(\hat{\Lambda}_{k0}(t))$ versus t, where the $\hat{\Lambda}$ are estimated cumulative intensities (Breslow, 1972, 1974), should reveal approximately parallel curves. This, however, can be hard to assess 'by eye'. Another graphical method is to plot $\hat{\Lambda}_{h0}(t), h = 1, \ldots, k-1$, against $\hat{\Lambda}_{k0}(t)$ for all or some selected values of t. Under the proportional hazards assumption, these curves should be approximately straight lines with intercepts zero (Andersen et al., 1993, p.539-545).

A different way to check the proportional hazards assumption is to use the score process and to derive tests based on properties of this process, which is the main purpose of this work. The score process, i.e. the partial derivative of the partial log likelihood function, is defined in the next section. Both existing tests and tests proposed here will be considered. If the score process cannot be accepted as a Brownian bridge, the null hypothesis of proportional hazards has to be rejected. A Kolmogorov-Smirnov type statistic to test this hypothesis has been proposed earlier (Therneau, Grambsch and Fleming, 1990). In this paper we propose and investigate the behavior of two other test statistics which are based on Anderson-Darling and Cramér-von Mises type statistics, respectively. Plotting the score processes in addition to the formal tests can also be useful.

The proposed proportional hazards tests will be discussed in Section 2. Simulation studies of the level and power of the tests, together with a simulated example, are presented in Section 3, and Section 4 gives a real data example. Finally, some concluding comments are given in Section 5.

2 Tests for Proportional Hazards Based on the Score Process

For the semi-parametric model (3), the partial likelihood is (Andersen and Gill, 1982)

$$PL = \prod_{i=1}^{n} \prod_{j=1}^{N_i(\tau_i)} \frac{e^{\boldsymbol{\beta}^T \mathbf{x}_i}}{\sum_{k \in R(t_{ij})} e^{\boldsymbol{\beta}^T \mathbf{x}_k}}, \tag{6}$$

where n is the number of processes, $\mathbf{x}_i$ is the covariate vector of process number i and $R(t_{ij})$ is the risk set at time t_{ij}, i.e. the set of units which can have an event and which are under observation at time t_{ij} . The τ_i and $N_i(\tau_i)$ are described in the introduction, Section 1. The maximum likelihood estimate for $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, is the $\boldsymbol{\beta}$ that maximizes this likelihood function. It can be shown that $\hat{\boldsymbol{\beta}}$ has an asymptotic normality property and also that it is consistent under certain rather general conditions (Andersen and Gill, 1982) . The variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ may be estimated consistently by $\mathcal{I}(\hat{\boldsymbol{\beta}})^{-1} = (n^{-1}\mathcal{I}_F(\hat{\boldsymbol{\beta}}))^{-1}$, where $\mathcal{I}_F(\boldsymbol{\beta})$ is the Fisher Information matrix (Andersen et al., 1993, p.486). In this estimate the censored observations, or processes without events, contribute by being a part of the risk set. The log partial likelihood is given by

$$PLL = \log PL = \sum_{i=1}^{n} \sum_{j=1}^{N_i(\tau_i)} \left\{ \boldsymbol{\beta}^T \mathbf{x}_i - \log \left(\sum_{k \in R(t_{ij})} e^{\boldsymbol{\beta}^T \mathbf{x}_k} \right) \right\} \tag{7}$$

The derivative of the PLL with respect to β_l, $l = 1, \ldots, p$, becomes

$$\begin{aligned} \left[\frac{\partial PLL}{\partial \beta_l} \right] &= \sum_{i=1}^{n} \sum_{j=1}^{N_i(\tau_i)} \left\{ x_{il} - \frac{\sum_{k \in R(t_{ij})} e^{\boldsymbol{\beta}^T \mathbf{x}_k} x_{kl}}{\sum_{k \in R(t_{ij})} e^{\boldsymbol{\beta}^T \mathbf{x}_k}} \right\} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{N_i(\tau_i)} \{ x_{il} - \bar{x}_l(\boldsymbol{\beta}, t_{ij}) \} \end{aligned} \tag{8}$$

where

$$\bar{x}_l(\boldsymbol{\beta}, t_{ij}) = \frac{\sum_{k \in R(t_{ij})} e^{\boldsymbol{\beta}^T \mathbf{x}_k} x_{kl}}{\sum_{k \in R(t_{ij})} e^{\boldsymbol{\beta}^T \mathbf{x}_k}} \tag{9}$$

is a weighted mean of the covariates over the risk set at time t_{ij}. Equation (8) can also be written on integral form as (Therneau et al., 1990)

$$\begin{aligned} \left[\frac{\partial PLL}{\partial \beta_l} \right] &= \sum_{i=1}^{n} \int_0^{\infty} \{ x_{il} - \bar{x}_l(\boldsymbol{\beta}, s) \} \, dN_i(s) \\ &\equiv U_l(\boldsymbol{\beta}, \infty) \end{aligned} \tag{10}$$

The process $U_l(\hat{\boldsymbol{\beta}}, t) = \sum_{i=1}^n \int_0^t \left\{ x_{il} - \bar{x}_l(\hat{\boldsymbol{\beta}}, s) \right\} dN_i(s)$, for $t > 0$, will be called the score process.

We shall now see how one may test the proportional hazards assumption, i.e. that the parameter $\boldsymbol{\beta}$ in model (3) is independent of time, by deriving test statistics based on the score process. Suppose that the covariate $\mathbf{x} = (x_1, \ldots, x_p)$ is independent of time, and we want to test the proportional hazards assumption for covariate x_l, i.e. the lth element in the covariate vector. Suppose further that x_l is (approximately) orthogonal to the other covariates. When the proportional hazards assumption holds we then have that (Therneau et al., 1990)

$$\sqrt{\widehat{Var}(\hat{\beta}_l)} U_l(\hat{\boldsymbol{\beta}}, \cdot) \xrightarrow{L} W^0 \left(\frac{\sigma_{ll}(\cdot)}{\sigma_{ll}(\infty)} \right), \tag{11}$$

where $W^0(\cdot)$ is a Brownian bridge (Billingsley, 1968). The quantity $\sigma_{ll}(\cdot)$ will be discussed later. Recall that the estimated variance of the parameter estimator is $\widehat{Var}(\hat{\beta}_l) = \{\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})\}_{ll}$.

Hence, under the null hypothesis of proportional hazards, $U_l(\hat{\boldsymbol{\beta}}, \cdot)$ will converge to a time-transformed Brownian bridge. This implies that if $U_l(\hat{\boldsymbol{\beta}}, t)$ is far from zero it is an indication that the null hypothesis is wrong. There are several ways to measure such discrepancies. Besides, it can be useful to plot $U_l(\hat{\boldsymbol{\beta}}, t)$ as well, to get an idea about how large the discrepancies are and whether they have a special pattern. Therneau et al. (1990) proposed to use the test statistic

$$KS = \sup_t \left| \sqrt{\widehat{Var}(\hat{\beta}_l)} U_l(\hat{\boldsymbol{\beta}}, t) \right| \xrightarrow{L} \sup_{u \in [0,1]} |W^0(u)|, \tag{12}$$

which is a Kolmogorov-Smirnov type statistic (Billingsley, 1968), to test the proportional hazards assumption. The computation of the KS statistic is straightforward in for instance S-Plus, using the standard output from Cox-regression.

Let us now return to the quantity $\sigma_{ll}(\cdot)$ in (11). Let

$$\begin{aligned}
S^{(0)}(\boldsymbol{\beta}, t) &= \frac{1}{n} \sum_{i \in R(t)} e^{\boldsymbol{\beta}^T \mathbf{x}_i} \\
S^{(1)}(\boldsymbol{\beta}, t) &= \frac{1}{n} \sum_{i \in R(t)} \mathbf{x}_i e^{\boldsymbol{\beta}^T \mathbf{x}_i} \\
S^{(2)}(\boldsymbol{\beta}, t) &= \frac{1}{n} \sum_{i \in R(t)} \mathbf{x}_i \mathbf{x}_i^T e^{\boldsymbol{\beta}^T \mathbf{x}_i}, \\
E(\boldsymbol{\beta}, t) &= S^{(1)}(\boldsymbol{\beta}, t) / S^{(0)}(\boldsymbol{\beta}, t) \\
\text{and} \;\; V(\boldsymbol{\beta}, t) &= S^{(2)}(\boldsymbol{\beta}, t) / S^{(0)}(\boldsymbol{\beta}, t) - E(\boldsymbol{\beta}, t) E(\boldsymbol{\beta}, t)^T
\end{aligned} \tag{13}$$

$E(\boldsymbol{\beta}, t)$ is in fact the quantity defined in (9) and $V(\boldsymbol{\beta}, t_{ij})$ is the covariance of $\boldsymbol{x}$ at time t_{ij}, defined analogously to the running mean, whose l, l' element is given by

$$V_{ll'}(\boldsymbol{\beta}, t_{ij}) = \frac{\sum_{k \in R(t_{ij})} e^{\boldsymbol{\beta}(t)^T \mathbf{x}_i} [x_{kl} - \bar{x}_l(\boldsymbol{\beta}, t_{ij})][x_{kl'} - \bar{x}_{l'}(\boldsymbol{\beta}, t_{ij})]}{\sum_{k \in R(t_{ij})} e^{\boldsymbol{\beta}(t)^T \mathbf{x}_i}} \tag{14}$$

Further we have that (Andersen et al., 1993, p.481-501)

$$\frac{1}{n}\mathcal{I}_F(\boldsymbol{\beta}, t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t V(\boldsymbol{\beta}, s) dN_i(s) = \int_0^t V(\boldsymbol{\beta}, s)\frac{d\bar{N}(s)}{n}, \tag{15}$$

where $\bar{N}(s) = \sum_{i=1}^n N_i(s)$. Asymptotically, we have that

$$\frac{1}{n}\mathcal{I}_F(\hat{\boldsymbol{\beta}}, t) \xrightarrow{P} V(t), \tag{16}$$

where

$$V(t) = \int_0^t v(\boldsymbol{\beta}_0, u) s^{(0)}(\boldsymbol{\beta}_0, u)\lambda_0(u) du, \tag{17}$$

is the asymptotic covariance matrix for $n^{-1/2}U(\boldsymbol{\beta}_0, \cdot)$. Here $\boldsymbol{s}^{(m)}(\boldsymbol{\beta}, t)$, $m = 0, 1, 2$, are assumed to be continuous functions of $\boldsymbol{\beta}$ uniformly in t, satisfying

$$\sup_{\boldsymbol{\beta}, t}\left\|\frac{1}{n}\boldsymbol{S}^{(m)}(\boldsymbol{\beta}, t) - \boldsymbol{s}^{(m)}(\boldsymbol{\beta}, t)\right\| \xrightarrow{P} 0 \text{ as } n \to \infty, \ m = 0, 1, 2, \tag{18}$$

$v(\boldsymbol{\beta}, t)$ is defined analogously to $V(\boldsymbol{\beta}, t)$ in (13), with s instead of S, and $\boldsymbol{\beta}_0$ is the true parameter (Andersen et al., 1993, p.497).

Now, $\sigma_{ll}(t) = V(t)_{ll}$ and it is assumed that $V(t)_{jk} = 0$ for $j \neq k$, which is the condition that requires x_l to be orthogonal to the other covariates (Therneau et al., 1990). Hence as an estimator we can use $\hat{\sigma}_{ll}(t) = \left\{\frac{1}{n}\mathcal{I}_F(\hat{\boldsymbol{\beta}}, t)\right\}_{ll}$.

We shall now define some alternative test statistics for the hypothesis of proportional hazards based on (11). Let $g_l(t) = \frac{\sigma_{ll}(t)}{\sigma_{ll}(\infty)}$, where $\sigma_{ll}(\infty) = \sigma_{ll}(\tau)$ for all practical purposes, where $\tau = \max \tau_i$, and τ_i is the endpoint of the observation interval of the ith process. Now (11) states that $\sqrt{\widehat{Var}(\hat{\beta}_l)}U_l(\hat{\boldsymbol{\beta}}, \cdot) \xrightarrow{L} W^0(g_l(\cdot))$.

As a first alternative to the KS statistic (12), we consider an Anderson-Darling type statistic. It follows from (11) that

$$AD \equiv \widehat{Var}(\hat{\beta}_l)\int_0^\tau \frac{U_l(\hat{\boldsymbol{\beta}}, t)^2}{g_l(t)(1 - g_l(t))} d(g_l(t)) \xrightarrow{L} \int_0^\tau \frac{W^0(g_l(t))^2}{g_l(t)(1 - g_l(t))} d(g_l(t)) = \int_0^1 \frac{W^0(u)^2}{u(1-u)} du, \tag{19}$$

where the right hand side has the Anderson-Darling distribution (Anderson and Darling, 1952).

Let us now turn to the computation of the test statistics. By introducing $0 < t_1^* \leq t_2^* \leq \cdots \leq t_N^* \leq \tau$ as the distinct event times, see Figure 1, and letting $D(t_k^*)$ be the set of processes which has an event at time t_k^*, we have that

$$U_l(\hat{\boldsymbol{\beta}}, t) = \sum_{i=1}^{n} \int_0^t [x_{il} - \bar{x}_l(\hat{\boldsymbol{\beta}}, s)] dN_i(s) = \sum_{t_k^* \leq t} \sum_{i \in D(t_k^*)} [x_{il} - \bar{x}_l(\hat{\boldsymbol{\beta}}, t_k^*)], \tag{20}$$

where $\bar{x}_l(\hat{\boldsymbol{\beta}}, s)$ is defined in (9). Note that $U_l(\hat{\boldsymbol{\beta}}, t) = 0$ for $t \leq t_1^*$ and for $t \geq t_N^*$ from the definition of $\hat{\boldsymbol{\beta}}$.

Let $u = g_l(t)$ and we get that

$$AD = \widehat{Var}(\hat{\beta}_l) \int_0^1 \frac{U_l(\hat{\boldsymbol{\beta}}, g_l^{-1}(u))^2}{u(1-u)} du \tag{21}$$

Since $U_l(\hat{\boldsymbol{\beta}}, g_l^{-1}(u))$ is a constant for $g_l^{-1}(u) \in [t_{i-1}^*, t_i^*]$, i.e. for $u \in [g(t_{i-1}^*), g(t_i^*)]$, we have that

$$\begin{aligned} AD &= \widehat{Var}(\hat{\beta}_l) \sum_{k=2}^{N-1} \int_{g_l(t_{k-1}^*)}^{g_l(t_k^*)} \frac{U_l(\hat{\boldsymbol{\beta}}, t_{k-1}^*)^2}{u(1-u)} du \\ &= \widehat{Var}(\hat{\beta}_l) \sum_{k=2}^{N-1} U_l(\hat{\boldsymbol{\beta}}, t_{k-1}^*)^2 \left[\ln\left(\frac{g_l(t_k^*)}{g_l(t_{k-1}^*)} \right) + \ln\left(\frac{1 - g_l(t_{k-1}^*)}{1 - g_l(t_k^*)} \right) \right] \\ &= \widehat{Var}(\hat{\beta}_l) \sum_{k=2}^{N-1} U_l(\hat{\boldsymbol{\beta}}, t_{k-1}^*)^2 \left[\ln\left(\frac{\sigma_{ll}(t_k^*)}{\sigma_{ll}(t_{k-1}^*)} \right) + \ln\left(\frac{\sigma_{ll}(\infty) - \sigma_{ll}(t_{k-1}^*)}{\sigma_{ll}(\infty) - \sigma_{ll}(t_k^*)} \right) \right], \end{aligned} \tag{22}$$

where $\sigma_{ll}(t)$ can be replaced by the consistent estimator $\hat{\sigma}_{ll}(t) = \left\{ \frac{1}{n} \mathcal{I}_F(\hat{\boldsymbol{\beta}}, t) \right\}_{ll}$. If the largest observation time is an event time, i.e. $t_N^* = \tau$, however, the sum should only be from 2 to $N-2$, otherwise we would get $\hat{\sigma}_{ll}(\infty) - \hat{\sigma}_{ll}(t_{N-1}^*) = \hat{\sigma}_{ll}(\tau) - \hat{\sigma}_{ll}(\tau) = 0$ in the denominator, since $V(\hat{\boldsymbol{\beta}}, t) = 0$ for $t > t_{N-1}^*$ in this case.

The second statistic considered is a Cramér - von Mises type statistic

$$CV \equiv \int_0^\tau \widehat{Var}(\hat{\beta}_l) U_l(\hat{\boldsymbol{\beta}}, t)^2 d(g_l(t)) \xrightarrow{L} \int_0^1 W^0(u)^2 du, \tag{23}$$

where the right hand side has the known Cramér - von Mises distribution (Cramér, 1928, von Mises ,1931). Note that CV can be computed in exactly the same way as described above for AD. We get

$$CV = \widehat{Var}(\hat{\beta}_l) \sum_{k=2}^{N-1} U_l(\hat{\boldsymbol{\beta}}, t_{k-1}^*)^2 \frac{(\hat{\sigma}_{ll}(t_k^*) - \hat{\sigma}_{ll}(t_{k-1}^*))}{\hat{\sigma}_{ll}(\infty)} \tag{24}$$

The CV statstic can be viewed as a measure of the (unsigned) area between the score process and the zero line. The weight function $\frac{1}{t(1-t)}$ in the AD statistic has the effect of giving greater importance to observations in the tails (Anderson and Darling, 1952). One motivation for introducing these test statistics, is that they should have power to detect both monotone and non-monotone alternatives.

As mentioned, all these test statistics have known asymptotic distributions. Using these (asymptotic) distributions, the null hypothesis of proportional hazards is rejected, on a 5% level, if $KS \geq 1.36$, $CV \geq 0.461$ or $AD \geq 2.492$ respectively.

The Kolmogorov-Smirnov statistic is supposed to be quite sensitive to alternatives for which covariates have a monotonically increasing or decreasing effect over time, specifically for alternatives such that $\lambda(t|X = z)/\lambda(t|X = y)$ is strictly monotone increasing in t for all $z < y$, or is strictly monotone decreasing for all $z < y$ (Therneau et al., 1990). The statistics AD and CV are supposed to perform well also when covariates have a non-monotonic effect on the intensity. This will be investigated further in the simulation study in Section 3.

In practice it is a bit complicated and very computer intensive to compute $\hat{\sigma}_{ll}(t)$ in for instance S-Plus in the case with recurrent events. There are ways to find approximations to these statistics, however. In particular when $\tau = \tau_0$ for all processes in a recurrent event case (as in the cardiac arrest data set which we study in Section 4), then $V(\boldsymbol{\beta}, t)$ is constant for all t since the risk set is constant. The expression for $V(t)$ in (17) therefore suggests to use $c\Lambda_0(t) = c\int_0^t \lambda_0(u)du$ as an estimate for $\sigma_{ll}(t)$, and we get for example

$$AD_{\Lambda_0} = \widehat{Var}(\hat{\beta}_l) \sum_{k=2}^{N-1} U_l(\hat{\boldsymbol{\beta}}, t^*_{k-1})^2 \left[\ln\left(\frac{\hat{\Lambda}_0(t^*_k)}{\hat{\Lambda}_0(t^*_{k-1})} \right) + \ln\left(\frac{\hat{\Lambda}_0(\infty) - \hat{\Lambda}_0(t^*_{k-1})}{\hat{\Lambda}_0(\infty) - \hat{\Lambda}_0(t^*_k)} \right) \right], \qquad (25)$$

and equivalently for CV_{Λ_0}. The quantity $\widehat{Var}(\hat{\beta}_l)$ is estimated as before.

In cases without recurrent events, S-Plus computes the estimator $\hat{\sigma}_{ll}(t) = \{\frac{1}{n}\mathcal{I}_F(\hat{\boldsymbol{\beta}}, t)\}_{ll}$ for all event times in a reasonably efficient manner.

3 A Study of the Tests of Proportional Hazards

Data sets from known models are simulated to study the behavior of the tests (see Reiersølmoen, 2000, for a more thorough discussion). S-Plus is used in the following studies.

Simulation studies were done to check the level and the power of the AD, CV and KS tests proposed in Section 2. For comparison, we also included the chi-square test based on the Schoenfeld residuals (Grambsch and Therneau, 1994) mentioned in the introduction. Only the test with the identity transform, $g(t) = t$, is studied here and the corresponding chi-square statistic is denoted T. Data sets simulated from a Cox model were used to study the the level of the tests and data sets simulated from models other than Cox models (where the proportional hazards assumption does not hold) were used to study the power of the different tests.

The computation of $\hat{\sigma}_{ll}(t) = \{\frac{1}{n}\mathcal{I}_F(\hat{\boldsymbol{\beta}}, t)\}_{ll}$ becomes very complicated and computer intensive in S-Plus for data sets with processes with recurrent events or more than one covariate. Thus data sets without recurrent events and only one covariate are used in the main study of the test statistics.

First, data sets are simulated from the case where $\lambda(t) = \sqrt{t}e^x$, $X \sim$ unif(0,1) and $\tau \sim$ unif(0,2.5). The percentage of processes with no events in these simulations are around 20-40 %. A Cox-model $\lambda(t) = \lambda_0(t)e^{\beta x}$ is fitted and the AD, CV, KS and T test statistics are calculated. Table 1 shows how many percent of the times the null hypothesis was rejected when this type of data set was simulated and the test statistics were calculated for 2000 data sets, and a 5 % nominal significance level was used. The first row shows the number of processes, n, simulated in each data set.

Table 1: *Rejection level for the test statistics AD, CV, KS and T (in %) for data simulated from the model with $\lambda(t) = \sqrt{t}e^x$, where each process has only one possible event. Each simulation is done 2000 times and n denotes the number of processes in each simulated data set.*

n	50	100	200	500	1000
AD	5.9	5.6	5.4	5.1	5.2
CV	5.7	5.3	5.1	5.2	5.2
KS	3.4	3.7	3.7	4.3	4.0
T	3.0	2.9	2.9	3.4	4.0

The AD and CV statistics seem to perform very similar and both achieve a level close to 5 %. The KS also performs quite well but seems to have a somewhat lower actual level than AD and CV and seems to be a somewhat too conservative test. The chi-square test performs similar to the KS test. Hence, it also seems to be a somewhat too conservative test. All test statistics seem to perform relatively equal for smaller as well as for larger data sets, but slightly better for larger data sets.

A typical plot of the score process, $U_1(\hat{\boldsymbol{\beta}}, t)$ (we have only one covariate), is shown in Figure 2 (a), together with two horizontal dotted lines which show the rejection limits of the KS test. This would be accepted as a Brownian bridge, it does not show any particular pattern or any large deviation from the zero line.

Then data sets from a case where $\lambda(t) = 2e^{5tx}$, $X \sim$ unif(0,1) and $\tau \sim$ unif(0.3,0.6) are simulated. In this case the covariate has a monotonically (increasing) effect over time. Again a Cox-model $\lambda(t) = \lambda_0(t)e^{\beta x}$ is fitted for each data set and the AD, CV, KS and T test statistics are calculated. Table 2 shows how many percent of the times the null

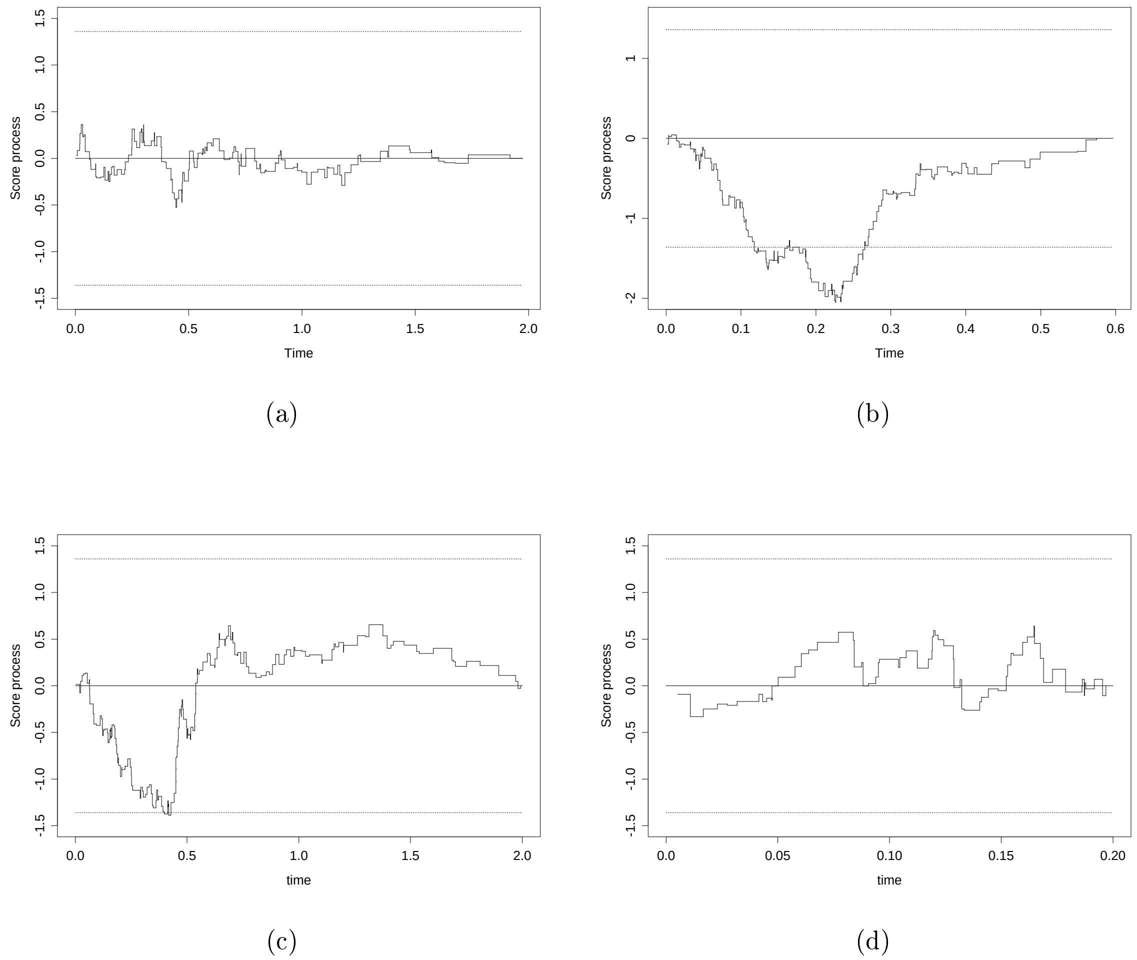

Figure 2: The score prosess $U_1(\hat{\beta}, t)$ plotted against time for covariate x, for data from the case where (a) $\lambda(t) = \sqrt{t}e^{x}$, (b) $\lambda(t) = 2e^{5tx}$, (c) $\lambda(t) = 2e^{\beta(t)x}$, where $\beta(t)$ is described in (26), and (d) $\lambda(t) = \sqrt{t}e^{2.8x}$ with recurrent events. The two horizontal dotted lines in each plot show the rejection limits for the KS test.

hypothesis was rejected when this type of data set was simulated and the test statistics were calculated for 1000 data sets, and a 5 % nominal significance level was used. The first row shows the number of processes, n, simulated in each data set.

Again, it seems like AD and CV performs very similar but AD has slightly higher power, especially for smaller data sets. The chi-square test statistic, T, seems to perform very similar to AD and CV for this type of data set. The KS test statistic seems inferior to AD and CV in this case. Yet for very large data sets all test statistics have large and almost equal power. For all types of test statistics the power is clearly better for larger sample

Table 2: *Rejection level for the test statistics AD, CV, KS and T (in %) for data simulated from the model with* $\lambda(t) = 2e^{5tx}$, *where each process has only one possible event . Each simulation is done 1000 times and n denotes the number of processes in each simulated data set.*

n	50	100	200	500	1000
AD	18.7	32.8	59.2	94.8	99.8
CV	18.3	31.2	57.4	94.3	99.9
KS	12.0	22.5	45.6	88.4	99.6
T	14.6	30.8	59.5	92.8	99.8

sizes. A typical pattern of the score process $U_1(\hat{\boldsymbol{\beta}}, t)$ is shown in Figure 2 (b). Here the process clearly indicates a deviation from a Brownian bridge, it is large negative for almost all time points. So this plot should be considered in addition to checking the test statistics to detect the nature of the deviations from the assumption of proportional hazards.

Next, data sets from the case where $\lambda(t) = 2e^{\beta(t)x}$, where

$$\beta(t) = \begin{cases} -\log(4) & 0 \leq t < 0.4 \\ \log(1) & 0.4 \leq t < 0.7 \\ -\log(4) & 0.7 \leq t \end{cases}, \tag{26}$$

$X \sim$ unif(0,1) and $\tau = 2$ were simulated. Now the effect of the covariate is varying non-monotonically with time. Again a Cox-model $\lambda(t) = \lambda_0(t)e^{\beta x}$ is fitted for each data set and the AD, CV, KS and T test statistics are calculated. Table 3 shows how many percent of the times the null hypothesis was rejected, on a 5 % level, when this type of data set was simulated and the test statistics were calculated 1000 times. The first row shows the

Table 3: *Rejection level for the test statistics AD, CV, KS and T (in %) for data simulated from the model with* $\lambda(t) = 2e^{\beta(t)x}$, *where* $\beta(t)$ *is described in (26) and each process has only one possible event. Each simulation is done 1000 times and n denotes the number of processes in each simulated data set.*

n	50	100	200	500	1000
AD	5.5	8.6	14.1	42.8	87.1
CV	6.1	8.0	13.3	41.1	87.5
KS	4.6	9.7	16.2	51.3	89.8
T	2.1	3.9	3.6	5.4	5.7

number of processes, n, simulated in each data set.

In this case, with a non-monotonic time dependency, the chi-square test does not seem to have any power at all. Somewhat surprisingly, KS seems to perform best of the test statistics in this case. Again, all types of statistics perform better for larger sample sizes, except the chi-square test. In this case the tests have low power for small sample sizes.

A typical pattern of the score process is shown in Figure 2 (c). Also here it is quite clear that the process is not a Brownian bridge. From this plot it can be seen that the effect of the covariate is non-monotonic as well, since we first have almost only negative values and then almost only positive values. So even if the null hypothesis is not rejected on the basis of the test statistics, the plot should be investigated and studied further if it seems like the process deviates much from the zero line.

Recurrent Event Cases

For the cardiac arrest data (see next section) we have recurrent events where all processes are censored at the same time ($\tau = 24$). As mentioned in Section 2 it is expected that the approximation with $c\hat{\Lambda}_0(t)$ for $\hat{\sigma}_{ll}(t)$ will work quite well in this situation, since the risk set is constant over time. It would be of interest to investigate this further by a simulation study.

Since it is very computer intensive to perform simulation studies of the level and power of the tests in the recurrent events situation, only the case with $\lambda(t) = \sqrt{t}e^{2.8x}$, $x \sim$ unif(0,1), a fixed censoring time $\tau = 0.2$ and $n = 200$ is studied to check the level of the test. The power has not been investigated further. The resulting censoring percentage, i.e. the percentage of the processes which has zero events, in this case is approximately 70-80 %, which is similar to the censoring percentage in the 'cardiac arrest data set'. For both AD_{Λ_0} and CV_{Λ_0} the simulations resulted in a level of 4.8 %, and for KS the level turned out to be 3.6 %. So in this situation it seems like it is fine to use the approximation with $c\hat{\Lambda}_0(t)$ as an estimator for $\sigma_{ll}(t)$. Also here the KS test statistic has a lower actual level than 5 %. Hence, we assume that the use of AD_{Λ_0} and CV_{Λ_0} is adequate in situations with recurrent events and where all the processes are censored at the same time. A typical pattern of the score process for this type of data is shown in Figure 2 (d). There is no evidence that this is not a Brownian bridge.

Let us now consider a single recurrent events example with non-constant censoring time. A data set is simulated from the model $\lambda(t|x) = e^{xt}$, with $n = 200$, $\tau \sim \text{unif}(0.8, 1.4)$ and $x \sim \text{unif}(0, 1)$. A Cox-model $\lambda(t|x) = \lambda_0(t)e^{\beta x}$ is fitted. The tests based on the score process do not lead to rejection of the null hypothesis of proportional hazards. The resulting test statistics are $KS = 1.18$, $CV_{\Lambda_0} = 0.35$ and $AD_{\Lambda_0} = 1.96$, while the 5% critical values are 1.36, 0.461 and 2.492 respectively. From a plot of the score process, shown in Figure 3, however, we would be suspicious about the proportionality since it only takes on negative values and does not seem like a Brownian bridge.

In Figure 4 the Schoenfeld residuals are plotted against time. There is an increasing

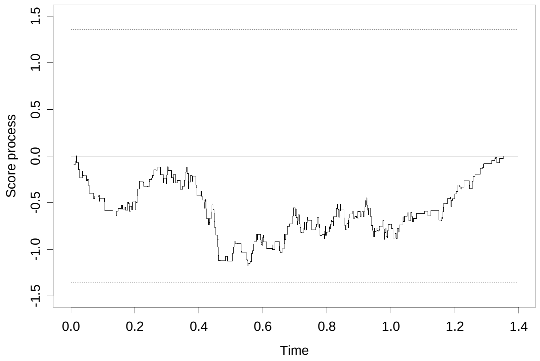

Figure 3: Plot of the score process, U_1, against time for the recurrent event data set simulated from the model $\lambda(t|x) = e^{xt}$, with $n = 200$.

tendency in this line, and the p-value of the chi-square test, T, is 0.028, which also indicates a departure from the assumption of time independent covariate effect.

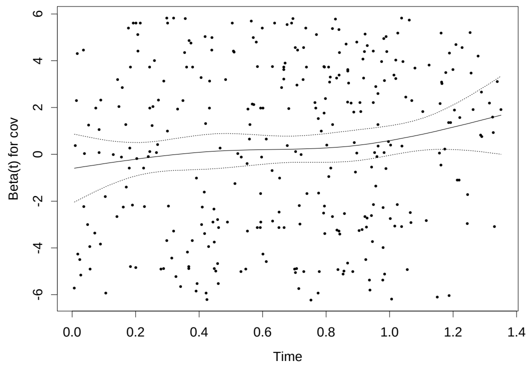

Figure 4: Schoenfeld residuals from model $\lambda_0(t)e^{\beta x}$, with a smoothed line with confidence bands plotted against time, for the recurrent event data set simulated from the model $\lambda(t|x) = e^{xt}$, with $n = 200$. (The covariate is denoted 'cov' in the text on the vertical axis.)

Figure 5 shows the log-minus-log plot (in the upper left corner) and plots of $\Lambda_{h0}(t)$ versus $\Lambda_{k0}(t)$, $h, k = 1, 2, 3$, i.e. we divided the data into three strata of equal size. The lines in the log-minus-log plot are not quite parallel, they do even seem to cross. Moreover, in the plots of $\Lambda_{h0}(t)$ versus $\Lambda_{k0}(t)$ the lines in the lower part of the figure do not seem totally straight. Hence, we would maybe suspect a departure from the assumed model. This is in accordance with the conclusion from the Schoenfeld residuals. It seems, however, easier to detect the tendencies of non-proportionality in Figures 3 and 4, particularly in Figure 3.

Since the graphical methods all seem to more or less clearly indicate a departure from the

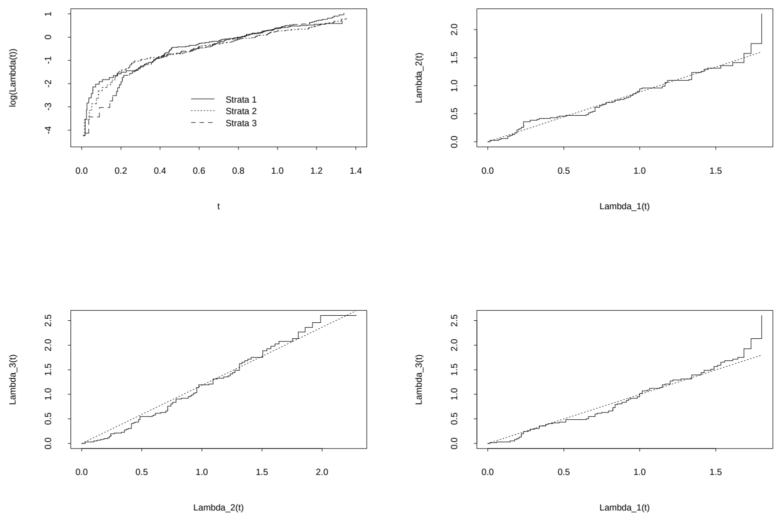

Figure 5: Log-minus-log plot and plots of $\Lambda_{h0}(t)$ versus $\Lambda_{k0}(t)$, from model $\lambda_h(t|x) = \lambda_{h0}(t)$, $h = 1, 2, 3$, stratified on x, for the recurrent event data set simulated from the model $\lambda(t|x) = e^{xt}$, with $n = 200$.

proportional hazards model, the results from the tests based on the score process may be somewhat disappointing. It is, however, important to remember that the CV_{Λ_0} and AD_{Λ_0} statistics are expceted to perform well only when all the processes have the same censoring time, and that is not the case here. According to simulations not reported we expect the CV_{Λ_0} and AD_{Λ_0} to be too conservative in the present example.

Discussion of the Simulation Study

According to the study of the level of the tests, the AD and CV tests perform almost equal and have the correct level. They seem slightly better than KS and the chi-square test which are too conservative. In the recurrent event case when all processes are terminated at the same time the AD_{Λ_0} and CV_{Λ_0} tests also have an actual 5 % level. The levels of the tests do not seem to depend much on the sample size.

When it comes to the power of the tests, the AD and CV tests perform better than the KS test in the example where the covariate has a monotonic effect over time, in particular for smaller sample sizes. The chi-square test in this case has approximately the same performance as AD and CV. In the example with a non-monotonic covariate effect over time studied here, however, the KS test seems in fact to perform a bit better than AD and CV, while the chi-square test has a total lack of power in this case.

Hence, the conclusion of this limited simulation study is that for practical use, the AD, CV and KS tests seem to be fairly equal, with the AD and CV tests having maybe slightly better overall properties. While the chi-square test should be used with care because of the total lack of power against non-monotonic alternatives.

In the real data example of the next section we will compute both AD_{Λ_0}, CV_{Λ_0}, KS and T, and compare the results. We will probably get the same conclusions from the different tests in most cases but together with a plot of the score process, this can give a better insight into the nature of the data.

4 The Cardiac Arrest Data Set

Description of the Data

In this section we analyze the cardiac arrest data set described in Kvaløy (1999), Irgens (2000) and Jakobsen (2000) by the methods discussed in the previous sections. This data set actually consists of a combination of two data sets from the years 1990 to 1998. The first set contains times of out-of-hospital cardiac arrests treated by the Emergency Medical Service located at the University Regional Hospital of Trondheim. These data for 1990 to 1994 were first studied by Skogvoll and Lindqvist (1999). See Skogvoll and Lindqvist (1999) for a detailed description of these data. The second part of the data contains weather observations recorded at the official weather station of the Norwegian Meteorological Institute at Værnes. These two data sets are combined and used to examine possible relationships between the occurrence of cardiac arrests and certain weather variables.

The weather variables considered in this study are air temperature, air pressure, relative humidity, wind speed, wind direction, precipitation and snow depth. A covariate 'snowfall' is also included, which is the increase in snow depth from day to day if this is positive and zero otherwise. Day number (1 to 365) is also included, as a covariate as an indication of year variation.

Air temperature is measured in °C, the pressure in Pa, relative humidity naturally in %, wind speed in m/s, precipitation in mm and snow depth and snow fall in cm. The wind direction is measured in radians. Wind from the north is represented by 0, wind from the east by $\pi/2$, south by π and wind from the west by $3\pi/2$. Since this is a 'circular' covariate and the effect of 0 and 2π should be the same, we have used a wind direction covariate calculated as (wind direction $- \pi)^2$, in the analysis. We could also have used another 'symmetry axis' but it is hard to decide which one is the best. Day number is also a 'circular' covariate, day 1= day 366, so we have chosen to use a day covariate, (day number $- 182)^2$, in the analysis.

For temperature, pressure, relative humidity, wind direction, wind speed and precipitation both one day and seven days averages, the average of the weather recordings at the day of the cardiac arrest and the six previous days, are used in the analysis. It is not clear if long term or short term effects are the most important. It is also possible to use weighted averages or averages over longer or shorter intervals. This is, however, not done in this study.

Amongst the patients with cardiac arrests there are 72.5 % men and 27.5 % women, and most of them are between 60 and 80 years old.

The Model

The time from midnight to midnight is regarded as a natural observation interval. This leads to a counting process, censored after 24 hours, which counts the occurrences of cardiac arrests during the day. The conditional intensity of this counting process is modeled by a Cox-model, $\lambda(t|\mathbf{x}) = \lambda_0(t)e^{\boldsymbol{\beta}^T\mathbf{x}}$. Here the weather variables are entries in the vector $\mathbf{x}$, and we want to find which of these are important for the intensity. In order to use the results of the previous sections we also need to make the reasonable assumption that given the covariates, $\mathbf{x}$, the events occur independently of each other, both within each counting process and between counting processes.

In the process of selecting the most important covariates we used stepwise regression based on a maximum likelihood criterion, together with some subjective reasoning. The Akaikes Information Criterion (AIC) and likelihood ratio tests were also considered. From this analysis the model found to describe the data best was the model with day covariate, snow fall and wind direction covariate using a one day average. The resulting model can then be written as

$$\lambda(t|\mathbf{x}) = \lambda_0(t)e^{1.48\cdot10^{-5}d+6.48\cdot10^{-2}s+3.79\cdot10^{-2}w}, \tag{27}$$

where d = day covariate, s = snow fall and w = wind direction covariate as one day average. The p-values for the coefficients are $2.8\cdot10^{-5}$, $2.1\cdot10^{-4}$ and $3.2\cdot10^{-2}$, respectively. Model (27) indicates that the intensity increases with increasing value of each of the covariates in the model. The intensity according to the model is highest in winter, when there is a heavy snow fall, and when the wind blows from the north.

The baseline intensity $\lambda_0(t)$ found by using a kernel estimator (the biweight kernel, Klein and Moeschberger (1997)), is shown in Figure 6. Since the intensity in this case is 'circular', i.e. the intensity is the same just before and just after midnight, shifted copies of the observations on the t-axis are replicated and then the kernel smoothing is applied to get the intensity curve to start and stop in the same point. This baseline intensity very much reflects people's day rhythm. At night when people are asleep the intensity is low. The moment people get up in the morning the intensity increases and is high all through the day when people are more active.

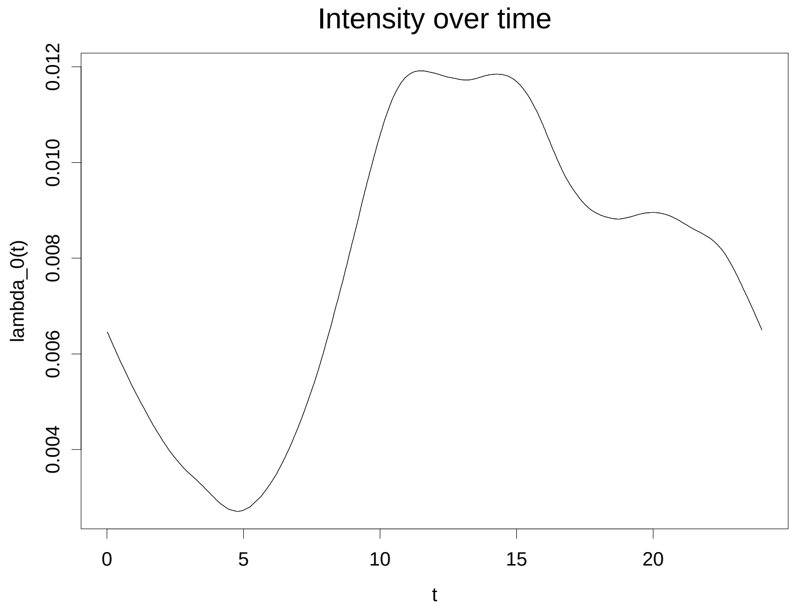

Figure 6: Estimate for the baseline intensity, $\hat{\lambda}_0(t)$, plotted against time for the cardiac arrest data set.

Checks of the Proportional Hazards Assumption

As mentioned in Section 1, there are several methods to check the proportional hazards assumption. We will here concentrate on the tests based on the score process, and in addition report the p-value from the chi-square test based on the Schoenfeld residuals for comparison, but the other checks should also be performed in analysing real life data to compare the results and get a better understanding of the data (see Reiersølmoen, 2000, for a more detailed analysis).

Let us consider one variable at the time. The score process $U_d(\hat{\boldsymbol{\beta}}, t)$ for the day covariate is shown in Figure 7. Even if it has a kind of a peak around 4 p.m., it can probably be accepted as a Brownian bridge. This is also confirmed from all the test statistics on a 5 % level, $AD_{\Lambda_0} = 0.92 < 2.492$, $CV_{\Lambda_0} = 0.20 < 0.461$ and $KS = 1.09 < 1.36$. Recall from Section 3 that AD_{Λ_0} and CV_{Λ_0} achieve (approximately) the correct 5 % level for data of this type. The p-value of 0.51 from the chi-square test based on the Schoenfeld residuals also indicate that that the proportional hazards assumption holds for the day covariate.

We now turn to the snowfall covariate. First we study the score process, $U_s(\hat{\boldsymbol{\beta}}, t)$ which is plotted in Figure 8. Again, this can probably be accepted as a Brownian bridge, which also the test statistics confirm, on a 5 % level, $AD_{\Lambda_0} = 0.61 < 2.492$, $CV_{\Lambda_0} = 0.11 < 0.461$ and $KS = 0.76 < 1.36$. Hence from these tests again, we conclude that the proportional hazards assumption holds. From the p-value of 0.96 from the chi-square test it also seems like the proportional hazards assumption holds for snow fall.

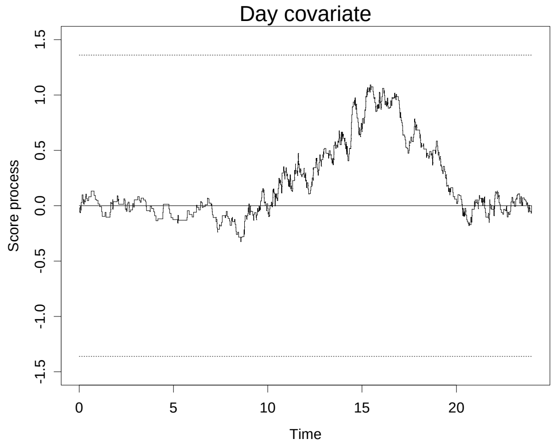

Figure 7: Plot of the score process, $U_d(\hat{\boldsymbol{\beta}}, t)$, for the day covariate against time for the cardiac arrest data set.

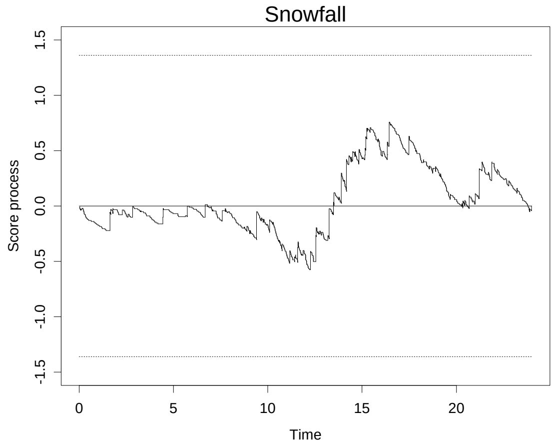

Figure 8: Plot of the score process, $U_s(\hat{\boldsymbol{\beta}}, t)$, for snowfall against time for the cardiac arrest data set.

Finally, we consider the wind direction covariate. Again we study the score process, $U_w(\hat{\boldsymbol{\beta}}, t)$, which is plotted in Figure 9 and as the test statistics confirm, this can be accepted as a Brownian bridge. The test statistics are $AD_{\Lambda_0} = 1.18 < 2.492$, $CV_{\Lambda_0} = 0.22 < 0.461$ and $KS = 0.91 < 1.36$. So again it seems like, on a 5 % level, that the proportional hazards assumption holds. The p-value of 0.57 from the chi-square test also seems to confirm this. Hence, it seems like the proportional hazards assumption holds for the wind direction covariate as well .

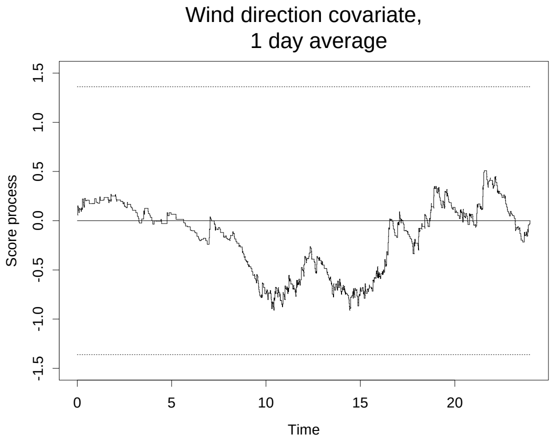

Figure 9: Plot of the score process, U_w, for the wind direction covariate, against time for the cardiac arrest data set.

Discussion of the Cardiac Arrest Data

From all the studies of the cardiac arrest data (Kvaløy, 1999, Irgens, 2000, Jakobsen, 2000) it seems clear that there is a seasonal effect on the intensity of cardiac arrests, the intensity is higher in winter, when it is more snow and lower temperatures. In some models, this effect is included as a day or week number covariate. In other models it is included as outdoor air temperature or snowfall. All these covariates represent much of the same. Similar results have been found from other studies as well. For example Nicolau et al. (1991) found seasonal differences in cardiac mortality and Enquselassie et al. (1993) found a relationsship between temperature and coronary diseases.

The present study verifies that the proportional hazards (or proportional intensities) assumption holds for the selected Cox-model used to model the cardiac arrest data. Further studies of goodness of fit in Reiersølmoen (2000) confirm that other aspects of the Cox-model like functional form of the covariate effect, the NHPP assumption and the overall fit of the model seem to be fulfilled, indicating that the Cox-model used is a reasonable model for the data.

5 Conclusion

There exist several methods to investigate the proportional hazards (or proportional intensities) assumption in the Cox model. We have here proposed some new tests based on the score process (AD and CV) and compared these tests to some existing tests (KS based on the score process and the chi-square test based on the Schoenfeld residuals). From the simulations it seems like the chi-square test based on the Schoenfeld residuals and the tests based on the score process perform fairly similar, except in situations where the covariates have a non-monotonic effect on the intensity. In such cases it may happen that the chi-square test has no power at all. The proposed tests, AD and CV, together with the KS test, may hence be better as overall tests, since they have power to detect non-monotonic alternatives as well. In the limited simulation study conducted here, the AD and CV tests are the most powerful tests against the monotonic alternative studied, while the KS test is slightly more powerful against the non-monotonic alternative. The AD and CV tests seem to have the best level properties in all cases considered. It is also recommended to plot the score process. It is generally hard to tell 'by eye' if a process is a Brownian bridge or not, but the plot of the score process gives useful information about tendencies in the data which supplements the information from the tests.

As mentioned before, the tests proposed here should be investigated further. For instance should both level and power of all the tests be studied for cases with recurrent events and cases with several covariates.

Acknowledgements

We would like to thank Bo Lindqvist for helpful discussions and suggestions during the work with this paper. Also thanks to Eirik Skogvoll for providing the cardiac arrest data. This work is from a part of Reiersølmoen's diploma thesis conducted during the spring 2000.

References

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

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study, *Annals* **10**: 1100–1120.

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain goodness of fit criteria based on stochastic processes, *Annals of Mathematical Statistics* **23**: 193–212.

Billingsley, P. (1968). *Convergence of Probability Measures*, John Wiley & Sons, New York.

Breslow, N. E. (1972). Contribution to the discussion of "Regression models and life-tables" by D. R. Cox, *Journal of the Royal Statistical Society, Series B* **34**: 187–220.

Breslow, N. E. (1974). Covariance analysis of censored survival data, *Biometrics* **30**: 89–99.

Cox, D. R. (1972). Regression models and life-tables, *Journal of the Royal Statistical Society, Series B* **34**: 187–220.

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

Cramér, H. C. (1928). On the composition of elementary errors, *Skand. Aktuar.* **11**: 13–74,141–180.

Enquselassie, F., Dobson, A. J., Alexander, H. M. and Steele, P. L. (1993). Season, temperature and coronary disease, *International Journal of Epidemiology* **22**: 632–636.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnosstics based on weighted residuals, *Biometrika* **81**: 515–526.

Irgens, P. (2000). *Statistical analysis of cardiac arrest data using parametric and non-parametric Poisson regression*, Master's thesis, Norwegian University of Science and Technology, Trondheim, Norway.

Jakobsen, T. (2000). *Statistisk analyse av hjertestansdata ved bruk av Cox-regresjon*, Master's thesis, Norwegian University of Science and Technology, Trondheim, Norway.

Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*, Springer-Verlag, New York.

Kvaløy, J. T. (1999). *Statistical Methods for Detecting and Modeling General Patterns and Relationships in Lifetime Data*, PhD thesis, Norwegian University of Science and Technology, Trondheim, Norway.

Nicolau, G. Y., Haus, E., Popescu, M., Sackett-Lundeen, L. and Petrescu, E. (1991). Circadian, weekly, and seasonal variations in cardiac mortality, blood pressure, and catecholamine excretion, *Chronobiology International* **8**: 149–159.

Reiersølmoen, L. (2000). *Goodness of fit methods in Cox's regression model for counting process data*, Master's thesis, Norwegian University of Science and Technology, Trondheim, Norway.

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

Skogvoll, E. and Lindqvist, B. H. (1999). Modeling the occurrence of cardiac arrest as a Poisson process, *Annals of Emergency Medicine* **33**: 409–417.

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

von Mises, N. (1931). *Wharscheinlichkeitsrechnung*, Deuticke, Leipzig.