Covariate order tests for covariate effect

Jan Terje Kvaløy
Department of Mathematical Sciences
Norwegian University of Science and Technology
N-7034 Trondheim, Norway
E-mail: jtk@stat.ntnu.no

February 18, 1999

Abstract

A new approach for constructing tests for covariate effect, i.e. association between a random right censored life time variable and a covariate, is proposed. The basic idea is to first arrange the observations in increasing order of the covariate and then base the test for covariate effect on a certain point process defined by the observation times.

Tests constructed by this approach are robust against outliers in the covariate values or misspecification of the covariate scale since they are only based on the ordering of the covariate. Of particular interest among the tests that can be constructed by this approach, is a test based on the Anderson-Darling statistic. Unlike most other tests for covariate effect, this test has power both against monotonic and nonmonotonic alternatives. A simulation study shows that this test has favorable properties compared to existing tests, both with regard to power and robustness.

The usefulness of applying robust tests based on the ordering of the covariate rather than tests which are using the observed values of the covariate is demonstrated by simulations as well as practical examples. Some approaches for testing for covariate effect in cases with several covariates are also briefly discussed.

KEY WORDS: Test for association; Bootstrap test; Permutation test; Robustness; Level properties; Power properties; Censored data

1 INTRODUCTION

In analysis of life time data a question of basic concern is if there is an association between the life time variable and possible explanatory variables. Statistical tests can be used to make inference about such relationships. This paper will mainly deal with the situation with only one candidate for explanatory variable, but the situation with several possible explanatory variables will also be briefly discussed.

A general formulation of the basic problem studied in this paper is that we want to test if there is an association between the two random variables Z and X where the first variable may be subject to right censoring. For convenience we will use the life time data terminology and call T the *observation time*, where $T = \min(Z, C)$ is the minimum of a *life time* Z and a *censoring time* C. The censoring time C is assumed to be conditionally independent of the life time Z given the *covariate* X. The life time Z is assumed to come from a distribution with finite first order moment and support on the positive real line. It is of course also assumed that $P(Z < C) > 0$. The covariate X can in general be time-dependent, and several of the tests discussed in this paper allow time-dependent covariates. We shall, however, focus on the situation with covariates which are constant in time, and unless otherwise stated, X is assumed to be constant in time. Based on n independent observations $(T_1, \delta_1, X_1), \ldots, (T_n, \delta_n, X_n)$ of (T, δ, X), where $\delta = I(Z \leq C)$, we want to test the null hypothesis of no *covariate effect*, or in other words, the null hypothesis that Z is *independent* of X.

Tests for covariate effect have been discussed by a number of authors. Jones (1991) study tests for covariate effect in a general class of tests for survival data problems proposed by Jones and Crowley (1989). This class of tests include for instance the Cox score test (Cox, 1972), the logit rank test of O'Brien (1978) further studied in O'Quigley and Prentice (1991), the Brown, Jr., Hollander and Korwar (1974) modification of the Kendall rank test for survival data and the generalized log-rank test (Jones and Crowley, 1989). Jones (1991) also suggested some new tests for covariate effect generated from the Jones & Crowley class of test statistics, in particular a modified generalized log-rank test. The tests for covariate effect in the Jones & Crowley class of tests are generally constructed for the alternative of a relative risk model. Other tests for covariate effect are for instance constructed for Aalen's linear model (Aalen, 1980), see also Aalen (1989) and Grønnesby (1997). All the above cited tests have the limitation of being constructed for the alternative of a monotonic covariate effect. Tests constructed for nonmonotonic alternatives are suggested by Le and Grambsch (1994) and McKeague, Nikabadze and Sun (1995). The former approach has the limitation of leading to tests with very low power against monotonic alternatives, while the latter has the limitation of requiring very large samples.

A new approach for constructing tests for covariate effect called the *covariate order* approach is presented in this paper. The idea is arising from the derivation of a method for exponential regression presented in Kvaløy and Lindqvist (1998a). The basic idea is to arrange the observations $(T_1, \delta_1, X_1), \ldots, (T_n, \delta_n, X_n)$ such that $X_1 \leq X_2 \leq \cdots \leq X_n$, and then create a certain point process based on the observation times $T_1, \ldots, T_n$. If the life times are exponentially distributed and there is no covariate effect, this process will be a homogeneous Poisson process (HPP). Thus tests of HPP versus not HPP applied to this process can be used as tests of covariate effect. By resampling techniques, this test approach can be generalized to other life time distributions than the exponential, requiring the added assumption that the censoring time C is independent of X.

Tests derived by the covariate order approach are robust against outliers in the covariate space since they are only using the ordering of the observed covariate values. Covariate order tests can in principle be used both for continuous and discrete covariates, and both for constant and time dependent covariates, though the approach is best suited for continuous covariates which are constant in time. Of particular interest is a test based on the Anderson-Darling statistic. This test has favorable properties as a test for covariate effect since it has good power both against monotonic and nonmonotonic alternatives, and is among the tests studied in this paper the recommended test for general use.

The covariate order approach for constructing tests for covariate effect is presented in Section 2, and other tests for covariate effect are reviewed in Section 3. A simulation study comparing small sample properties of a number of tests for covariate effect is presented in Section 4, and some real data examples are presented Section 5. The problem of testing for covariate effect in situations with several covariates is briefly considered in Section 6, and finally some concluding comments are given in Section 7.

2 THE COVARIATE ORDER TEST APPROACH

In Section 2.1 we consider the case when Z, conditionally given X, is exponentially distributed with hazard rate $\lambda(X)$. For this case, the covariate order approach for testing the null hypothesis of no covariate effect, that is $\lambda(X) \equiv \lambda$, is presented. Then in Section 2.2 this approach is generalized to allowing Z to have any life time distribution. Maximum likelihood tests are also briefly mentioned.

The covariate order test approach is not suited for taking into account the effect of time-dependent covariates. Thus, the covariate X is assumed to be constant in time, or, if not, represented by its value at time 0.

2.1 Exponentially Distributed Life Times

Kvaløy and Lindqvist (1998a) presented a method for exponential regression which for the case of one single continuous covariate was based on ordering the observations in increasing order of the covariate. The same idea can be used to construct tests for covariate effect. Arrange the observations $(T_1, \delta_1, X_1), \ldots, (T_n, \delta_n, X_n)$ such that $X_1 \leq X_2 \leq \ldots \leq X_n$. Observations with equal covariate values are arranged in random order. Then let the observation times $T_1, \ldots, T_n$, irrespectively if they are censored or not, be subsequent inter-arrival times of an artificial point process on a "time" axis s. For this process, let points which are endpoints of intervals corresponding to *uncensored* observations be considered as events, occurring at times denoted $S_1, S_2, \ldots$. This is visualized in Figure 1, for an example where the ordered observations are $(T_1, \delta_1 = 1), (T_2, \delta_2 = 0), (T_3, \delta_3 = 1), \ldots, (T_{n-1}, \delta_{n-1} = 0), (T_n, \delta_n = 1)$. Generally $S_i = \sum_{j=1}^{k(i)} T_j$ where $k(i) = \min\{r | \sum_{j=1}^{r} \delta_j = i\}$.

If there is no covariate effect, $\lambda(x) \equiv \lambda$, then the process $S_1, \ldots, S_{k(n)}$ is a HPP. Otherwise the process $S_1, \ldots, S_{k(n)}$ is not a HPP. This observation suggests using statistical tests constructed for testing the null hypothesis of a HPP versus various alternatives as tests for covariate effect. Tests for covariate effect based on the process $S_1, \ldots, S_{k(n)}$ will be called *covariate order tests*. It was argued in Kvaløy and Lindqvist (1998a) that the process $S_1, \ldots, S_{k(n)}$ can be viewed as being approximately a nonhomogeneous Poisson process (NHPP) if $\lambda(x)$ is varying reasonably smoothly as a function of x. This motivates adopting tests constructed for testing HPP versus NHPP as tests for covariate effect.

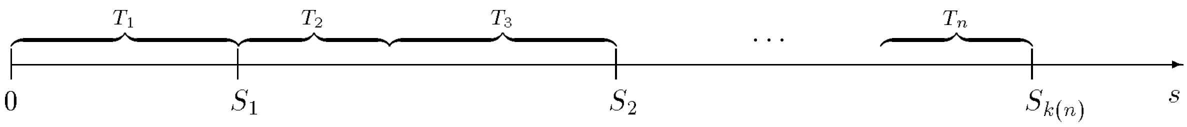

Figure 1: Construction of the process $S_1, \ldots, S_{k(n)}$.

There exist a number of tests in the literature for testing HPP versus NHPP, see for instance Ascher and Feingold (1984) or Kvaløy and Lindqvist (1998b) and references therein. One of the most popular and frequently used tests in the HPP versus NHPP setting is the Laplace test. Let $S = \sum_{i=1}^{n} T_i$ and for convenience define

$$\hat{k} = \begin{cases} k(n) & \text{if } S_{k(n)} < S \\ k(n) - 1 & \text{if } S_{k(n)} = S \end{cases}$$

Then the test statistic of the Laplace test is

$$LAP = \frac{\sum_{i=1}^{\hat{k}} S_i/S - \hat{k}/2}{\sqrt{\hat{k}/12}}.$$

This statistic is approximately normally distributed under the null hypothesis, and the approximation is very good even for very small samples. The Laplace test has optimal properties against certain monotonic alternatives, see Bain, Engelhardt and Wright (1985) or Cox and Lewis (1966).

A problem, however, with the Laplace test and many other tests for HPP versus NHPP is lack of power against nonmonotonic alternatives. If $\lambda(x)$ is a monotonic function of x, the intensity of the process $S_1, \ldots, S_{k(n)}$ will be a monotonic function of s, and using a test like the Laplace test will be appropriate. In practice, however, $\lambda(x)$ may well be nonmonotonic implying that the intensity of the process $S_1, \ldots, S_{k(n)}$ is nonmonotonic.

In Kvaløy and Lindqvist (1998b) a test of HPP versus NHPP based on the Anderson-Darling statistic was studied. This test, called the Anderson-Darling test for trend, has the favorable property of having power both against monotonic and nonmonotonic alternatives. See Kvaløy and Lindqvist (1998b) for details. The test statistic for the Anderson-Darling test for trend is

$$AD = -\frac{1}{\hat{k}} \left[\sum_{i=1}^{\hat{k}} (2i-1)(\ln \frac{S_i}{S} + \ln(1 - \frac{S_{\hat{k}+1-i}}{S})) \right] - \hat{k}$$

The asymptotic null distribution of this statistic was derived by Anderson and Darling (1952) and an explicit expression for the limiting cumulative distribution was given in Anderson and Darling (1954). Even for sample sizes smaller than 10 the asymptotic distribution is a good approximation to the exact distribution. On a 5% level the null hypothesis is rejected if $AD \geq AD_{0.05} = 2.492$. The Anderson-Darling test for trend used as a test for covariate effect as explained above will be called the Anderson-Darling test for covariate effect.

Notice that no assumption of continuity of X is needed for constructing the covariate order tests above. The test statistics will still have the correct distribution under the null hypothesis and can still be used if X for example is a discrete covariate. Covariate order tests can not in general be expected to work well in cases with a time dependent covariate unless the ordering of the covariate remain approximately the same over time, as only the information of the values of the covariate at time 0 is used.

Tests for covariate effect in exponentially distributed data can also be constructed in other ways than suggested above. An obvious approach is to construct likelihood ratio tests based on some parametric model of $\lambda(x)$. Such tests will have good power properties against alternatives similar to the parametric model used, but not generally against other alternatives, and will not necessarily be robust against outliers in the covariate values and misspecifications of the covariate scale. In the simulation study in Section 4 the likelihood ratio test for testing if $\beta = 0$ in the parametric model $\lambda(x) = \exp(\alpha + \beta x)$ is included.

2.2 General Life Time Distributions - Resample Tests

Using the test approach described in the previous subsection in the case of non-exponentially distributed life times may be seriously misleading since the null hypothesis of HPP is then violated. Lindqvist, Kjønstad and Meland (1994) survey a set of possible fallacies by using tests constructed for the HPP versus NHPP setting for testing the more general null hypothesis of a renewal process. Elvebakk (1998) shows that these fallacies can be overcome by using resampling techniques. Similar resampling approaches, respectively a bootstrap and a permutation approach, can be adapted in the present situation as well.

The resampling test approach presented in this subsection is, unlike the other tests for covariate effect discussed in this paper, based on testing the null hypothesis that (T, δ) is independent of X. If this null hypothesis is rejected, we want to conclude that Z is depending on X. For this to be the case we need to add to the assumptions stated in the introduction the extra assumption that the censoring time C is *independent* of the covariate X. In other words, random censoring is assumed. It is also tacitly assumed that the null hypothesis of independence between (T, δ) and X implies that Z and X are independent.

Let the test statistic of a test for covariate effect constructed for exponentially distributed life times be denoted TE (in the previous subsection TE is LAP, AD or the test statistic of the likelihood ratio test). For simplicity we assume that the the null hypothesis is rejected for large values of TE (if the Laplace test is used let $TE = |LAP|$).

The bootstrap test approach starts by splitting the original observations $(T_1, \delta_1, X_1), \ldots, (T_n, \delta_n, X_n)$ in two parts, one consisting of the pairs $(T_1, \delta_1), \ldots, (T_n, \delta_n)$ and one consisting of just $X_1, \ldots, X_n$. Then draw n new observation triples $(T_1^*, \delta_1^*, X_1^*), \ldots, (T_n^*, \delta_n^*, X_n^*)$ where the (T^*, δ^*) pairs are drawn randomly with replacement from $(T_1, \delta_1), \ldots, (T_n, \delta_n)$ and the X^* are independently drawn randomly with replacement from $X_1, \ldots, X_n$. Then calculate TE^* based on the bootstrap sample $(T_1^*, \delta_1^*, X_1^*), \ldots, (T_n^*, \delta_n^*, X_n^*)$. Repeat this B times to obtain the sample $TE^{*(1)}, \ldots, TE^{*(B)}$. The empirical distribution of this sample is an approximation to the null distribution of the test statistic TE. Let TE_{obs} be the observed value of TE calculated from the original observations. An approximate p-value of the test is then given by $\hat{p} = \sum_{i=1}^{B} I(TE^{*(i)} > TE_{obs})/B$.

The permutation approach is similar to the bootstrap approach, except that the resampling is done *without* replacement. A permutation of the original sample can be written $(T_1, \delta_1, X_{\pi(1)}), \ldots, (T_n, \delta_n, X_{\pi(n)})$ where $\pi(1), \ldots, \pi(n)$ denotes a permutation of the numbers

$1, \ldots, n$. The exact permutation null distribution of TE is found by calculating $TE^{\star}$ for all $n!$ possible permutations π of the original observations. In practice it is sufficient to calculate $TE^{\star(1)}, \ldots, TE^{\star(P)}$ for a large number, P, of permutations, and calculating the approximate p-value $\hat{p} = \sum_{i=1}^{P} I(TE^{\star(i)} > TE_{obs})/P$.

The bootstrap test and the permutation test may seem to be very similar tests, and indeed, in practice they give similar results in most cases (see Section 4 and comments below). Conceptually, however, the two test approaches are quite different. The bootstrap approach is based on assuming that the empirical distributions of (T, δ) and X are reasonable approximations to the respective true distributions. By sampling independently from these two empirical distributions, an approximation to the null distribution of TE in the "bootstrap world" where the empirical distributions are the true distributions, are found. The exact null distribution of TE in this "bootstrap world" is found by letting $B \to \infty$. As $n \to \infty$ the "bootstrap world" approaches the real model. The bootstrap approach is intuitively appealing, and will under certain conditions lead to tests with asymptotically correct level (see for instance Beran 1986, 1988 or Davison and Hinkley 1997, p. 38). It is beyond the scope of this paper to formally verify that these conditions hold for the bootstrap tests studied here. Instead we rely on simulations for verifying that the bootstrap tests have desirable level properties. In fact it turns out, see Section 4, that the bootstrap tests work well even for fairly small samples.

The permutation test is a conditional combinatorial test. In contrast to the bootstrap test, the exact permutation test will have correct level for samples of any size. This can be verified as follows. Under the null hypothesis the statistic $D = \{(T_1, \delta_1), \ldots, (T_n, \delta_n), X_{(1)}, \ldots, X_{(n)}\}$, where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ are the ordered values of $X_1, \ldots X_n$, is statistically sufficient. Thus, the distribution of TE given $D = d$ is independent of the unknown distributions of (T, δ) and X. Then in principle the exact conditional p-value

$$p(d) = P(TE > TE_{obs} | D = d)$$

can be determined. The conditional distribution of TE given $D = d$ is the distribution giving probability $\frac{1}{n!}$ to each of the permutations π of $(T_1, \delta_1, X_{\pi(1)}), \ldots, (T_n, \delta_n, X_{\pi(n)})$. Hence, $p(d)$ can be found by calculating TE for all these permutations, or, in practice, approximated by calculating TE for a large number, P, of permutations.

For a general introduction to bootstrap and permutation tests see for instance Davison and Hinkley (1997, chap. 4). They comment that while the bootstrap approach apply to a much wider class of testing problems than the permutation approach, the two approaches give very similar results in situations where both apply. This fact has been stated more precisely by Romano (1989) for certain situations. Simulations commented in Section 4 show that the bootstrap and the permutation approach give very similar results in the present case as well.

For simulations choosing B or P equal to 1000 (or less) is in most cases sufficient (see Davison and Hinkley 1997, chap. 4.2.5). For calculating p-values for real data somewhat larger values are recommended if the p-value is small.

Resampling methods have here been presented as a way of generalizing tests for covariate effect constructed for the special case of exponentially distributed life times, but these methods do of course also, assuming random censoring, apply to other tests for covariate effect. For instance may resampling in some cases be used to improve the level properties of a test compared to when the asymptotic distribution of the test statistic is used. Examples of this are given in Section 4.

Whether the data strongly contradict the assumption of no covariate effect in the censoring distribution can be checked by using a resample test on the censored observations alone, or on the sample $(T_1, 1-\delta_1, X_1), \ldots, (T_n, 1-\delta_n, X_n)$ and assuming no covariate effect in the life time variable.

3 Other Tests

Jones and Crowley (1989) present a general class of nonparametric tests for survival data which unify a number of existing tests and generate new tests for the s-sample (comparing equality of $s \geq 2$ samples), s-sample trend and single continuous covariate problems. The asymptotic properties of this class is studied in Jones and Crowley (1990), and the single continuous covariate problem is further studied in Jones (1991). The tests for covariate effect in this class are appropriate tests for testing $\beta = 0$ the relative risk models with hazard rate of the form $\lambda(t|X(t)) = \lambda_0(t)h(\beta X(t))$ where $h(\cdot)$ is an arbitrary monotone non-negative differentiable function. Notice that the covariate is allowed to be time-dependent.

Let $K(t)$ be a function of the covariate $X(t)$ and let $q_n(t)$ be some weight function. Also define a risk indicator $Y_i(t)$ where $Y_i(t) = 1$ if item i is at risk at time t. Then the Jones & Crowley class of test statistics can be written

$$W(q, K) = n^{-1/2} \sum_{i=1}^{n} \delta_i q_n(T_i)[K_i(T_i) - \bar{K}(T_i)] \tag{1}$$

where $\bar{K}(t) = (\sum_{j=1}^{n} Y_j(t)K_j(t))/Y.(t)$ is the average of $K(t)$ of those at risk at time t. Here $Y.(t) = \sum_{i=1}^{n} Y_i(t)$. Let $m(t)$ be the number of failures at time t. Then an estimator of the variance of (1) is given by

$$V(q, K) = n^{-1} \sum_{i=1}^{n} \sum_{i=1}^{n} \delta_i q_n(T_i)^2 C_{KK}(T_i) \frac{Y.(T_i) - m(T_i)}{Y.(T_i) - 1}$$

where $C_{KK}(t) = Y.(t)^{-1} \sum_{j=1}^{n} Y_j(t)(K_j(t) - \bar{K}(t))^2$ is the sample variance of those at risk at time t. See Jones and Crowley (1989) and Jones (1991) for details. Under the null hypothesis of no covariate effect ($\beta = 0$), $W(q, K)/\sqrt{V(q, K)}$ is asymptotically standard normally distributed (Jones and Crowley, 1990).

There is a number of important special cases of the Jones & Crowley class of tests. Let $K(t) = X(t)$ and $q_n(t) = 1$, then $W(1, X)$ is equivalent to the Cox score test (Cox, 1972) for the situation with a single covariate. A problem with the Cox score test is that it is not robust against outliers in the covariate space due to its direct use of the covariate values. A more robust approach is to instead use the rank of the covariate values. Let $r_j(t)$ be the rank of $X_j(t)$ among those at risk at time t and let $R_j(t) = r_j(t)/(Y.(t)+1)$. Mid-ranks are used if several covariate values have the same rank. Then $W(1, R)$ is the generalized log-rank statistic (Jones and Crowley, 1989). Other special cases of the Jones & Crowley class of tests is the logit rank test of O'Brien (1978) further studied in O'Quigley and Prentice (1991), and the Brown, Jr. et al. (1974) modification of the Kendall rank test for survival data.

Jones (1991) study these and some other tests and suggest some new candidates for robust tests of covariate effect. The conclusion of his study was to recommend the Cox score test and a modification of the generalized log-rank test. The Cox score test performed best both in small and large sample comparisons as long as there were no outliers in the covariate values, while the modified generalized log-rank test was recommended based on robustness and good over all performance. The modified generalized log-rank test can be written $W(\hat{\sigma}_{RX}, R)$, where $\hat{\sigma}_{RX}(t)$ is the sample covariance between R and X for those at risk at time t. The sample covariance term is introduced to improve the efficiency of the test.

In the Jones & Crowley class of tests the alternative is generally assumed to be a relative risk model. An alternative model which may generate different tests is Aalen's linear model (Aalen, 1980). For the case of a single covariate Aalen's hazard rate model can be written $\lambda(t|X(t)) = \beta_0(t) + \beta_1(t)X(t)$ where $\beta_0(t)$ and $\beta_1(t)$ are unspecified possibly time dependent regression functions. For more than one covariate, additional terms of the same form are just added to the model. Tests for covariate effect in Aalen's linear model were discussed in Aalen (1980, 1989) and Grønnesby (1997). For the case of one covariate, the test statistic of a test proposed in Aalen (1989) and studied in Grønnesby (1997) can be written

$$ALM = \frac{\sum_{i=1}^{n} \delta_i (X_i(T_i) - \bar{X}(T_i))}{\sqrt{\sum_{i=1}^{n} \delta_i (X_i(T_i) - \bar{X}(T_i))^2}}$$

and is asymptotically standard normally distributed under the null hypothesis. Here $\bar{X}(t) = Y.(t)^{-1} \sum_{i=1}^{n} Y_i(t) X_i(t)$. Notice that the numerator is identical to the Cox score test statistic while the variance estimator is simpler. Grønnesby (1997) suggested improving this test by re-estimating the variance under the null hypothesis. For the case of one covariate, however, this yields the Cox score test.

A problem with all the tests for covariate effect discussed so far in this section is that they are constructed for the alternative of a monotonic covariate effect. McKeague et al. (1995) present a test that can detect arbitrary forms of dependence between the possibly censored life time variable and one single continuous covariate. The test statistic is obtained from a certain test statistic process indexed by time and covariate, which converges in distribution to a Brownian sheet. Despite its ability to detect general alternatives, this test has limited practical usefulness as the recommended sample size for use of the test is, according to McKeague et al. (1995), in general at least 500 observations, and at least 1000 observations under heavy censoring. Moreover, the test statistic is somewhat complicated to calculate and requires choice of covariate strata and a bandwidth.

Le and Grambsch (1994) suggest a simple modification of the generalized log-rank test that leads to a test with power against certain nonmonotonic covariate effects. They suggest that instead of ranking the covariate values in increasing order, the smallest covariate value should be given rank one, the biggest covariate value rank two, the second smallest rank three, the second biggest rank four and so on. This can be done in any rank based test procedure, and leads to tests with power against U-shaped covariate effects, but lack of power against other, and in particular monotonic, covariate effects.

Another approach for constructing tests with power both against monotonic and nonmonotonic covariate effects is as follows. Divide the covariate axis in p intervals and introduce the p indicator variables $I_1, \ldots, I_p$ where $I_i(X_j) = 1$ if X_j is in group i, and 0 otherwise. Then fitting the Cox-model

$$\lambda(t|x) = \lambda_0(t) \exp(\beta_1 I_1 + \cdots + \beta_{p-1} I_{p-1})$$

and using the Cox score test to test the null hypothesis that $\beta_1 = \ldots = \beta_{p-1} = 0$, leads to an test which should have power both against monotonic and nonmonotonic alternatives. Open questions are how to choose the intervals and how many intervals to use. Simulations further commented in Section 4.4 have indicated that a reasonable approach is to divide the covariate axis in three intervals such that 1/3 of the observations fall in each interval. This test is included in the simulation study.

4 Simulation Study

In this section some of the tests discussed in the previous sections are compared in a simulation study. The tests considered are the three tests for exponential regression models suggested in Section 2.1, resample versions of these tests, the Cox score test, the covariance adjusted generalized log-rank test, the test of covariate effect in Aalen's linear model and the test based on dividing the covariate axis in three intervals. The abbreviations AD for the Anderson-Darling test for covariate effect, LAP for the Laplace test, LR for the likelihood ratio test, COX for the Cox score test, CGL for the covariate adjusted generalized log-rank test (the test in Section 3 based on the test statistic $W(\hat{\sigma}_{RX}, R)$), ALM for the test in Aalen's linear model and COX3 for the test based on dividing the covariate axis in three intervals are introduced. The term *covariate-based* tests will be used for the LR, COX and ALM tests, which are using the observed values of the covariate, while the other tests which are only using the order of the covariate values will be denoted *rank-based* tests.

All tests are evaluated at a 5% nominal significance level. Rejection probabilities are estimated by generating 5000 samples. Letting p denote the true rejection probability, this implies that the standard deviation of the estimated rejection probability is $\sqrt{p(1-p)/5000} \leq 1/\sqrt{20000} \approx 0.007$. For all resample tests 1000 bootstrap or permutation samples are drawn. The simulations are done in C and S-PLUS.

4.1 Level

First some simulations which illustrate the level properties of the tests on small samples are reported. In these simulations resample versions of all tests are considered. In Table 1 the simulated level of the various tests are reported for samples of size 10, 30 and 50 generated from a model with hazard rate $\lambda(t|x) = 1$ and a uniform(0,2) censoring distribution corresponding to 43% censoring. Notice that the standard deviation of the estimated rejection probability in this case is approximately 0.003.

	$n = 10$			$n = 30$			$n = 50$		
Test	Ord.	Boot.	Perm.	Ord.	Boot.	Perm.	Ord.	Boot.	Perm.
AD	0.055	0.044	0.052	0.055	0.045	0.048	0.051	0.046	0.053
LAP	0.053	0.050	0.050	0.051	0.047	0.048	0.049	0.049	0.054
LR	0.073	0.050	0.053	0.060	0.051	0.051	0.055	0.049	0.052
COX	0.060	0.053	0.051	0.057	0.051	0.047	0.054	0.050	0.052
CGL	0.051	0.053	0.050	0.052	0.050	0.051	0.051	0.047	0.053
ALM	0.022	0.075	0.055	0.046	0.063	0.051	0.047	0.054	0.051
COX3	0.083	0.060	0.053	0.067	0.052	0.048	0.063	0.053	0.054

Table 1: Simulated rejection probability based on 5000 simulations of samples of size 10, 30 and 50 using the hazard rate $\lambda(t|x) = 1$ and a uniform(0,2) censoring distribution corresponding to 43% censoring. Both ordinary tests, bootstrap tests and permutation tests are reported. 1000 repetitions are used in the resample tests.

We see that not all of the ordinary tests achieve the correct level for small samples. The ALM test is too conservative, while in particular the COX3 test, but also the LR test and to some extent the COX test, are nonconservative. For increasing samples size the level properties of the tests improves. For $n = 50$ all tests except the COX3 test which still is a bit nonconservative, seems to achieve a reasonably correct level. As expected, the permutation tests always achieve the correct level. Bootstrapping succeed to correct the level of all tests except for the ALM test and the COX3 test for small samples. These are the two tests with worst level properties, and if one of these tests are to be used on a small sample the permutation version of the tests should be used. Otherwise it is indifferent with regard to level properties whether a bootstrap or a permutation test is used. Simulations not reported have shown that it is also of little practical difference with regard to power properties whether a bootstrap or a permutation test is used. In the rest of this simulation study only bootstrap resample tests are considered.

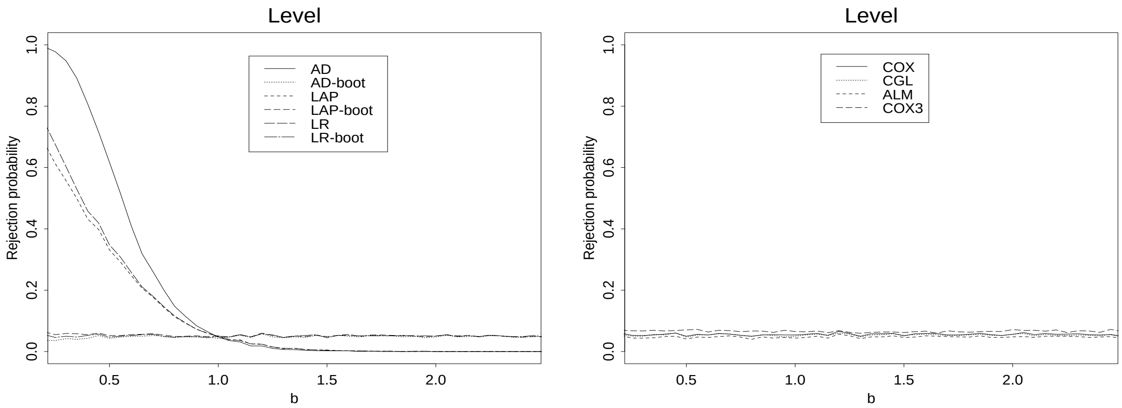

Figure 2: Estimated rejection probability as a function of b for samples of 50 observations from a model with hazard rate bt^{b-1} and no censoring. Lines are drawn between the estimates of the rejection probability for different b-values. 5000 repetitions for each b-value and a 5% significance level is used. 1000 repetitions are used in the bootstrap tests.

Another illustration of the level properties of the tests are given in Figure 2. In this example samples of size 50 are generated using the hazard rate $\lambda(t|x) = bt^{b-1}$ and no censoring. The rejection probability of the tests are simulated on a grid of different b-values with straight lines connecting the estimates at each b-value. We see that the AD, LAP and LR tests are very sensitive to deviations from exponentiality, while the bootstrapped AD, LAP and LR tests remain on the 5% level for all b-values. This illustrates that bootstrapping has the desired effect of generalizing the tests for covariate effect constructed for the exponential distribution to other life time distributions. Permutation tests have the same property. The COX, CGL and ALM tests also achieve approximately the correct level for all b-values, with the ALM test being slightly conservative. The COX3 test though is a bit nonconservative.

4.2 Power

Power properties of the various tests are illustrated in Figure 3 for models with hazard rate respectively $\lambda(t|x) = \exp(bx)$ and $\lambda(t|x) = \exp(bx^2)$, and with covariates drawn respectively from a uniform[0,1] and a uniform[-1,1] distribution. The censoring variable is drawn from the density $f_C(v) = 1.25v^{0.25}/V^{1.25}, \quad 0 \leq v \leq V$ where V is chosen to give approximately 50% censoring. Samples of size 50 are generated.

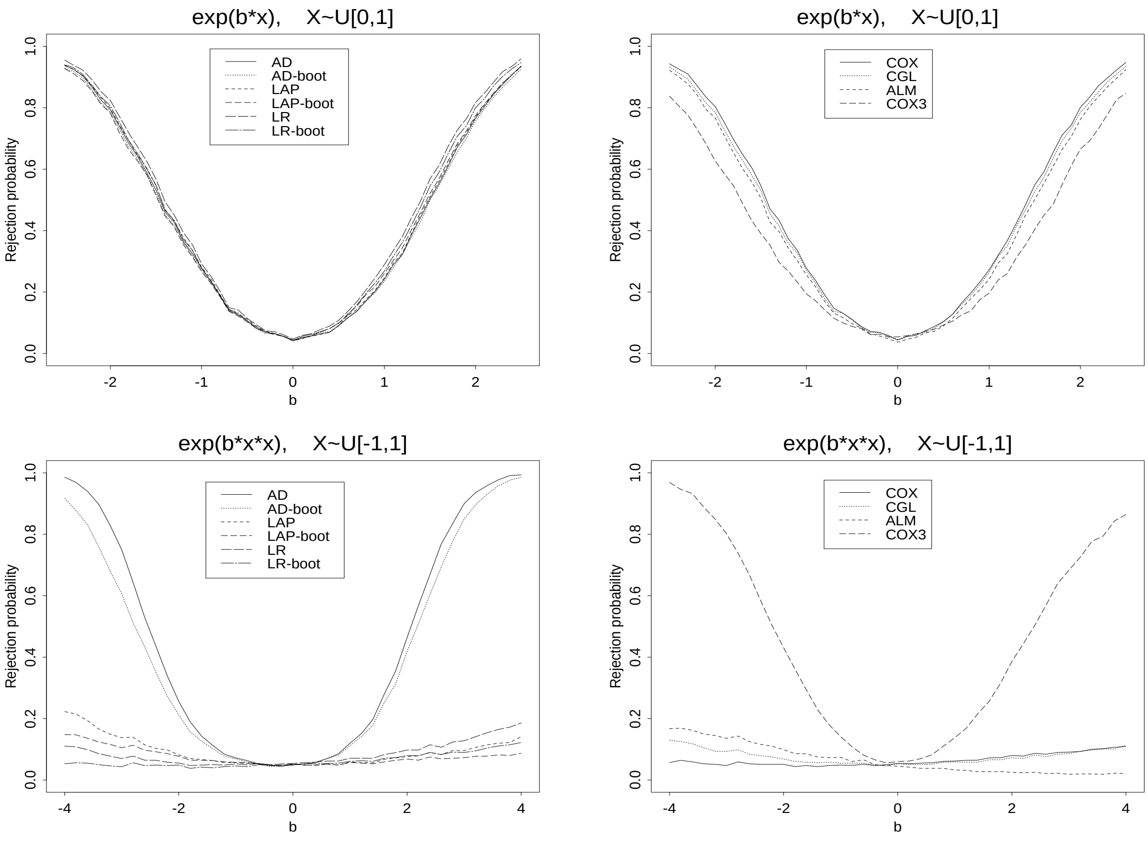

Figure 3: Estimated rejection probability as a function of b for samples of 50 observations from models with hazard rates $\exp(bx)$ and $\exp(bx^2)$, covariates drawn respectively from a uniform[0,1] and a uniform[-1,1] distribution and 50% censoring. Lines are drawn between estimates of the rejection probability for different b-values. 5000 repetitions for each b-value and a 5% significance level is used. 1000 repetitions are used in the bootstrap tests.

The first row in Figure 3 illustrates that all the tests, except the COX3 test, have approximately the same power properties against the monotonic covariate effect considered. The COX3 test is somewhat less powerful than the other tests even though it is slightly nonconservative. In particular note that the bootstrapped and the original versions of the AD, LAP and LR tests have approximately the same rejection probability for all b-values. This shows that the loss in power by bootstrapping the AD, LAP and LR tests is minimal. At the same time the gain in generality by bootstrapping the tests, from only being tests for covariate effect in the exponential distribution to tests for covariate effect in any life time distribution, is tremendous. Again, permutation tests have the same property. In the rest of the simulation study only the bootstrapped versions of the AD, LAP and LR tests are used.

The second row of Figure 3 illustrates the power properties of the tests in a case with nonmonotonic covariate effect. These plots illustrates that the tests constructed for monotonic alternatives have a total lack of power against this nonmonotonic alternative, while the AD test, the bootstrapped AD test and the COX3 test have good power properties against this nonmonotonic alternative as well.

Another example for comparing the power of the tests against both monotonic and nonmonotonic covariate effects is considered by simulating data from a model with hazard rate $\lambda(t|x) = 1.5t^{0.5}\exp(\cos(2\pi x))$. Samples of size 50 with no censoring are generated, and the covariate values are drawn uniformly over different intervals. These intervals and the corresponding simulated rejection probabilities are presented in Table 2. In the monotonic cases,

Test	U[0,0.25]	U[0,0.5]	U[0,0.75]	U[0,1]
AD-boot	0.508	0.992	0.905	0.641
LAP-boot	0.528	0.990	0.734	0.029
LR-boot	0.520	0.994	0.851	0.043
COX	0.535	0.994	0.847	0.042
CGL	0.510	0.990	0.785	0.044
ALM	0.515	0.988	0.706	0.010
COX3	0.395	0.974	0.904	0.829

Table 2: Simulated rejection probability based on 5000 simulations of samples of size 50 using the hazard rate $\lambda(t|x) = 1.5t^{0.5}\exp(\cos(2\pi x))$ no censoring and covariate values drawn from the uniform distributions indicated. 1000 repetitions are used in the bootstrap tests.

which are reported in the two first columns, all tests, except the COX3 which is less powerful, have fairly equal rejection probabilities. In the nonmonotonic cases in the two last columns, and in particular in the highly nonmonotonic case in the last column, the bootstrapped AD test and the COX3 test are more powerful than the other tests.

McKeague et al. (1995) compare the test suggested in their paper (see Section 3), a test which has power against general alternatives, to the COX test for samples of size 500 and 1000. These simulations show that their test has more power than the COX test against a nonmonotonic alternative, while the COX test has far more power against monotonic alternatives. The simulation study in McKeague et al. (1995) has been repeated here for the case of 500 observations. The censoring regime used in these simulations is to let the censoring variable be the minimum of 1 and an exponentially distributed variable with parameter chosen according to the degree of censoring wanted. The simulation results are presented in Table 3, and the hazard rates used are displayed in the table. The covariate values are drawn from a uniform[0,1] distribution. The test of McKeague et al. (1995) is denoted MNS and the rejection probabilities reported for this test are taken from McKeague et al. (1995). The results in Table 3 clearly shows that the bootstrapped AD test and the COX3 test are far better tests against general alternatives than the test suggested by McKeague et al. (1995).

Finally an example with a discrete covariate is given. In this example the covariate values are drawn uniformly among the three values 0, 1, 2. Samples of size 50 with no censoring are generated for different hazard rates $\lambda(t|x)$. The COX3 test is of course slightly modified in this situation. Instead of dividing the observations in three groups such that each group contains 1/3 of the observations, the observations are divided in three groups according to the three covariate values.

The estimated rejection probabilities for different hazard rates are reported in Table 4. The first column of this table shows that all tests still achieve the correct level except the CGL test which is a bit too conservative, and the COX3 test which again is nonconservative. Otherwise the table shows that there is surprisingly little difference between the covariate-based and the rank-based tests even in this case with discrete covariate. The covariate-based tests tends

Test	27% censoring e^x	e^{2x}	$7.5\min(x,1-x)$	60% censoring e^x	e^{2x}	$7.5\min(x,1-x)$
MNS*)	0.207	0.455	0.393	0.090	0.212	0.187
AD-boot	1.000	1.000	1.000	0.974	1.000	1.000
LAP-boot	1.000	1.000	0.128	0.976	1.000	0.073
LR-boot	1.000	1.000	0.056	0.978	1.000	0.037
COX	1.000	1.000	0.061	0.977	1.000	0.036
CGL	1.000	1.000	0.136	0.976	1.000	0.072
ALM	1.000	1.000	0.168	0.975	1.000	0.129
COX3	1.000	1.000	1.000	0.933	1.000	1.000

Table 3: Simulated rejection probability based on 5000 simulations of samples of size 500 using the hazard rates indicated and covariate values drawn from a uniform(0,1) distribution. 1000 repetitions are used in the bootstrap tests. *): The rejection probabilities of this test are taken from McKeague et al. (1995).

to be a bit more powerful than the rank-based tests against monotonic alternatives, but the difference is not big. In the cases with nonmonotonic covariate effects, the bootstrapped AD test and the COX3 test are far more powerful than the other tests. The COX3 is the most powerful test against the nonmonotonic alternatives, but is the least powerful test against monotonic alternatives, and is a nonconservative test.

Test	1	$1+x$	$1+0.8\cos(\pi x)$	$e^{x/2}$	$e^{2\cos(\pi x)}$	$0.1+t(x+1)$	$0.1+t\lvert x-1\rvert$
AD-boot	0.052	0.785	0.858	0.988	0.997	0.647	0.875
LAP-boot	0.050	0.785	0.005	0.958	0.005	0.663	0.006
LR-boot	0.048	0.857	0.055	0.996	0.450	0.725	0.045
COX	0.054	0.851	0.023	0.997	0.040	0.771	0.022
CGL	0.037	0.788	0.025	0.991	0.040	0.693	0.023
ALM	0.046	0.834	0.002	0.989	0.003	0.748	0.003
COX3	0.073	0.775	0.999	0.695	1.000	0.673	1.000

Table 4: Simulated rejection probability based on 5000 simulations of samples of size 50 using the hazard rates indicated, no censoring and covariate values drawn uniformly from {0,1,2}. 1000 repetitions are used in the bootstrap tests.

4.3 Robustness

So far we have assumed that the observed recorded covariate value X is equal to the true covariate value which can be denoted $X^\diamond$. For various reasons, for instance measurement errors, errors in records, misspecification of scale or other sources of *covariate contamination*, this need not always be the truth. For instance may outliers in the observed covariate values be due to such covariate contamination. Jones (1991) study the robustness of tests for covariate effect under different models for covariate contamination. The simulation study of Jones (1991) has been repeated here. In all simulations the hazard rate model considered is $\lambda(t|x) = 1+3x$, the censoring distribution is uniform[0,2] and samples of size 50 are generated. Three covariate distributions are considered, exponential(2), N(2,1) and uniform[0,2], and three types of covariate contamination are considered:

1. No contamination, $X = X^{\diamond}$.
2. 10% contamination, $X = X^{\diamond} + \gamma J$ where $\gamma = 3$ and $P(J = 1) = 1 - P(J = 0) = 0.1$ and J is independent of $X^{\diamond}$.
3. For the exponential covariate there is 10% contamination from an exponential(0.5) distribution, for the normal covariate there is 10% contamination from a N(5,1) distribution and for the uniform covariate there is 10% contamination from an uniform[2,5] distribution.
4. Misspecification of scale, $X = \exp(X^{\diamond})$.

The results are presented in Table 5

Covar. dist.	Type of contam.	AD-boot	LAP-boot	LR-boot	COX	CGL	ALM	COX3
exp(2)	1	0.777	0.777	0.823	0.863	0.795	0.687	0.651
	2	0.567	0.569	0.246	0.301	0.578	0.163	0.485
	3	0.643	0.654	0.372	0.449	0.661	0.256	0.528
	4	0.783	0.779	0.645	0.790	0.780	0.332	0.664
N(2,1)	1	0.914	0.634	0.925	0.790	0.769	0.751	0.664
	2	0.783	0.506	0.747	0.533	0.600	0.500	0.512
	3	0.788	0.505	0.621	0.366	0.588	0.305	0.510
	4	0.902	0.626	0.766	0.658	0.723	0.387	0.663
U[0,2]	1	0.842	0.856	0.858	0.846	0.828	0.837	0.704
	2	0.677	0.672	0.349	0.391	0.652	0.344	0.547
	3	0.679	0.672	0.309	0.352	0.647	0.283	0.546
	4	0.838	0.853	0.797	0.807	0.824	0.762	0.710

Table 5: Simulated rejection probability based on 5000 simulations of samples of size 50 using the hazard rate $\lambda(t|x) = 1 + 3x$ and a uniform[0,2] censoring distribution. The covariate distributions are indicated and the types of covariate contamination are defined in the text. 1000 repetitions are used in the bootstrap tests.

As expected, we see from Table 5 that the bootstrapped LR test, the COX test and the ALM test, which are the covariate-based tests, generally are less robust against covariate contamination than the rank-based tests. The bootstrapped AD test is in Table 5 the test with best overall power properties, while the bootstrapped LAP test and the CGL test also have good general power properties. The COX3 test is generally a bit less powerful than the other rank-based tests.

4.4 Comments

A number of new tests for covariate effect can be established by using the covariate order approach for constructing tests for covariate effect. Two examples of such tests, the resampled AD test and the resampled LAP test, have been compared to a selection of other tests for covariate effect in this simulation study. The conclusion is that the two tests, both with regard to power properties and robustness, are useful alternatives to the existing tests for covariate effect. In particular is this the case for the resampled AD test. In terms of power against general alternatives, correct level and robustness, this is the recommended test among the tests studied. If we only want or need power against monotonic alternatives, the resampled AD test is still a safe choice, but the resampled LAP test or the CGL test can be equally good choices for robust tests in such cases. If there is no need for robustness against outliers or misspecification of scale, the resampled LR test and the COX test are tests with generally at least as good power properties against monotonic covariate effects as the other tests. The ALM test is generally a bit less powerful than the other two covariate-based tests in the cases studied. It is interesting to observe that, even though the covariate-based tests in most cases are at least as powerful as the rank-based tests, the difference in power between the covariate-based tests and the rank-based tests is in general not large. This is so even though the latter tests only use information on the rank of the observed covariate values.

The COX3 test is in many cases the most powerful test against nonmonotonic alternatives, but is generally the least powerful test against monotonic alternatives. It is also a nonconservative test unless the sample size is large. Thus the power properties of the COX3 test can not really be compared to the other tests. For practical use on small samples a resample version of the COX3 test should be used instead of the original test to be sure that the test have correct level. Since the COX3 test only group the covariates in three groups this test is even more robust against covariate contamination than the other rank-based tests.

In simulations not reported tests similar to the COX3 test, obtained by dividing the covariate axis in $p > 3$ intervals with an equal number of observations in each interval, has been tried. It turned out that one of these tests, obtained by dividing the axis in $p = 4$ intervals is a useful alternative to the COX3 test in the cases studied. This test is a bit more powerful than the COX3 test against nonmonotonic alternatives, but a bit less powerful against monotonic alternatives and a bit more nonconservative. Dividing the covariate axis in $p > 4$ intervals gave tests which for the cases considered in this paper were increasingly nonconservative, less powerful against monotonic alternatives, and not necessarily more powerful against nonmonotonic alternatives.

Tests for covariate effect constructed by the covariate order approach, and rank-based tests in general, are best suited for cases with continuous covariates so that ties do not make a problem. However, it is demonstrated in this simulation study that such tests can successfully be used in situations with discrete covariates as well. Notice that if a covariate order test is used to test the significance of a discrete covariate, the outcome of the test will depend on the (random) order in which the observations with equal covariate values are arranged when calculating the test statistic. Thus, rather than basing the test on a test statistic calculated from a random ordering of the observations, the test could be based on the mean or median of the test statistic calculated for all or a large number of the possible orderings of the data. The null distribution of this mean or median is approximated by taking the same mean or median for all resamplings of the original data. The power properties of this test procedure has not yet been exploited.

5 EXAMPLES

5.1 Glioma Data

Jones and Crowley (1989) give an illustrative example by considering a data set of post-treatment survival times and ages at time of treatment for 28 male patients with low-grade gliomas (brain tumors). The data set presented in Jones and Crowley (1989) is given in Table 6.

Surv.	Age	Cens.	Surv.	Age	Cens.	Surv.	Age	Cens.	Surv.	Age	Cens.
6	38.4	1	474	38.1	1	797	36.9	0	1349	22.2	0
61	59.6	1	535	57.3	1	954	49.0	1	1506	42.0	0
179	18.9	0	547	34.5	0	967	26.8	0	1517	43.1	0
236	65.5	1	587	37.0	0	1050	66.9	1	1624	30.9	1
296	67.8	1	637	30.1	0	1134	30.4	1	1828	35.7	0
370	42.1	1	639	49.5	1	1141	41.6	1	1983	29.5	0
420	66.2	1	747	50.9	1	1213	35.1	0	2237	57.8	0

Table 6: Survival times (days), ages (years) and censoring indicator (0=censored, 1=uncensored) for 28 male patients with low-grade gliomas. The data are taken from Jones and Crowley (1989).

For illustrating the effect of extreme covariate values on the various test statistics, Jones and Crowley (1989) also considered a modified data set. For this modified data set, the age of the last patient is changed from 57.8 to 97.8. Perhaps an unlikely value, but clearly a value within the range of possible values. The test statistics and p-values of the various tests are reported in Table 7. Since we only have 28 observations, the COX3 test is bootstrapped to assure that we have a test with correct level. For the unmodified data all tests finds the

	Unmodified data		Modified data	
Test	Statistic	p-value	Statistic	p-value
AD-boot	3.53	0.0134	3.81	0.0109
LAP-boot	2.60	0.0060	2.56	0.0084
LR-boot	7.85	0.0061	1.11	0.3088
COX	3.15	0.0016	1.40	0.1615
CGL	3.16	0.0016	2.72	0.0065
ALM	2.48	0.0131	1.82	0.0688
COX3-boot	8.54	0.0209	8.54	0.0184

Table 7: Test statistics and p-values for the glioma data. 1000 repetitions are used for estimating the p-values of the bootstrap tests.

covariate effect of age to be significant on a 5% level. The COX test and the CGL tests yield the lowest p-values. For the modified data, however, only the rank based tests yield significant results. This illustrates the considerable influence of a single extreme covariate value on the covariate-based tests, whereas the rank-based tests remain robust. Notice that the test statistic of the bootstrapped COX3 test, which is the most robust test, remain unchanged from the unmodified to the modified data.

5.2 Leukemia Data

Feigl and Zelen (1965) present uncensored data on survival times for patients with leukemia. These data are also studied by Cox and Oakes (1984). We here only consider the effect of the covariate white blood cell count for all 33 patients. The results of using the various tests to test the significance of this covariate is reported in Table 8. The bootstrapped AD test

Test	Statistic	p-value
AD-boot	5.66	0.0099
LAP-boot	3.15	0.0068
LR-boot	5.05	0.0683
COX	2.10	0.0357
CGL	2.76	0.0058
ALM	1.78	0.0751
COX3-boot	8.58	0.0231

Table 8: Test statistics and p-values for testing the covariate effect of white blood cell count in the leukemia data. 1000 repetitions are used for estimating the p-values of the bootstrap tests.

the bootstrapped LAP test and the CGL test yield clearly significant results, the COX test and the bootstrapped COX3 test give p-values below 0.05 as well, while the bootstrapped LR test and the ALM test give p-values larger than 0.05. The difference in p-values between the rank-based and the covariate-based tests indicates that there may be outliers or extreme covariate values present. A histogram of the white blood cell counts is displayed in Figure 4. We see from this histogram that there are some observations located away from the main

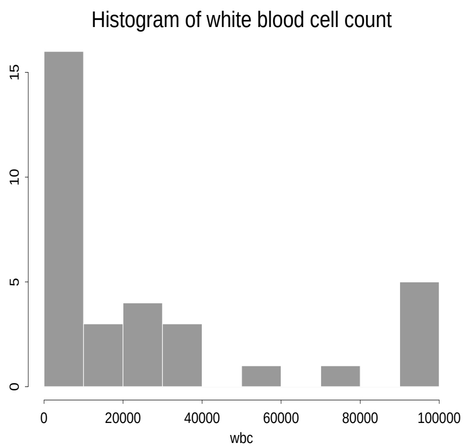

Figure 4: Histogram of white blood cell counts in the leukemia data.

cluster of observations. This probably causes the somewhat large p-values of some of the covariate based tests. A question is if the covariate has been measured on the correct scale.

The histogram in Figure 4 indicates that a logarithmic scale possibly is more appropriate. By using the logarithm of the white blood cell count as covariate, the results of the tests for covariate effect become as reported in Table 9. With the covariate taken on this scale, all

Test	Statistic	p-value
AD-boot	5.66	0.0104
LAP-boot	3.15	0.0064
LR-boot	10.33	0.0087
COX	3.10	0.0019
CGL	2.86	0.0042
ALM	2.96	0.0031
COX3-boot	8.58	0.0254

Table 9: Test statistics and p-values for testing the covariate effect of the logarithm of white blood cell count in the leukemia data. 1000 repetitions are used for estimating the p-values of the bootstrap tests.

the tests yield a significant result, and the COX test the most significant result. Notice that the test statistic of the CGL test is slightly different in the two cases since this test is not purely rank based. The histogram of the logarithm of the white blood cell count is displayed in Figure 5.

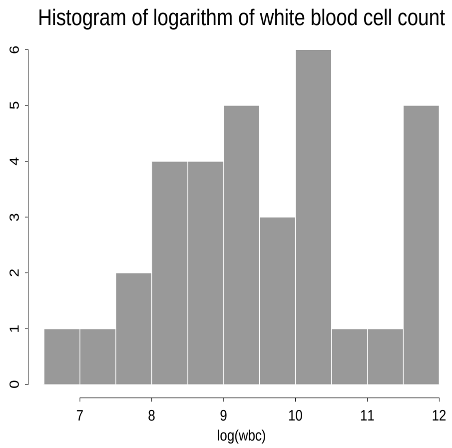

Figure 5: Histogram of the logarithm of white blood cell counts in the leukemia data.

This example illustrates the covariate-based tests sensitivity to the choice of scale, a choice which is not always obvious. This greatly benefits the rank-based tests for which the choice of scale is indifferent.

5.3 Cardiac Arrest Data

This final example examine data for out-of-hospital cardiac arrests reported to a Norwegian hospital over a 5 years period. The data have previously been presented and analyzed in Skogvoll and Lindqvist (1999). A conclusion of that paper is that cardiac arrests occur as a Poisson process.

In the present example the possible relationship between the frequency of cardiac arrests and certain weather covariates are examined. The weather observations are recorded three times a day at a local weather station, and the recording closest in time to each reported cardiac arrest is used as corresponding observed covariate values. The time variable is the time between consecutive cardiac arrests. None of the observations are censored, and the total data set consists of 524 observations. In this example only the bootstrapped AD test, is used.

Covariate	p-value
Air temperature	0.016
Air pressure	0.207
Relative humidity	0.624
Wind speed	0.967
Wind direction	0.171

Table 10: Estimated p-values for the effect of various weather covariates using the bootstrapped AD test with 10000 repetitions.

In Table 10 the significance of various weather covariates possibly related to the frequency of cardiac arrests are reported. It turns out that the only weather covariate with significant influence on the time between consecutive cardiac arrests is air temperature.

A further analysis of the data is done by dividing the data in certain relevant patient group subsets, and examine if the frequency of cardiac arrests in these groups relates significantly to air temperature. The results are reported in Table 11, and seems intuitively reasonable. It is reasonable that the frequency of cardiac arrests of non-cardiac origin is not related to

Subset of data	p-value	n
All cardiac arrests	0.016	524
Outdoors	0.011	196
At home	0.053	328
Cardiac origin	0.009	449
Non-cardiac origin	0.417	75
Age $>= 60$	0.002	366
Age < 60	0.167	157

Table 11: Estimated p-values for the effect of outdoor air temperature in certain subsets of the cardiac arrest data. The number of observations in each subset, n, is also reported. The subset denoted non-cardiac origin consist of cardiac arrests caused by external factors such as drowning, intoxication and so on. The estimated p-values are found by using the bootstrapped AD test with 10000 repetitions.

temperature, while the observation that the frequency of cardiac arrests only is significantly related to temperature for patients of age greater than 60 can be explained physiologically by the difference in nature of the typical cardiac arrest in each patient group. For a further study of these data see Kvaløy and Lindqvist (1998a) and Skogvoll and Lindqvist (1999).

6 SEVERAL COVARIATES

So far only testing the covariate effect of one single covariate, without regard to the possible presence of other covariates, has been considered. In practice, however, the case is often that there are several covariates possibly associated with the life time variable. For simplicity only the situation with covariates which are constant in time are discussed in this section, but for instance the tests based on the classical Cox-model or Aalen's linear model immediately generalizes to time-dependent covariates.

We first consider how to test the null hypothesis that there is no association between Z and a vector of covariates $\boldsymbol{X}$, versus the alternative that there might be an association between Z and one or several components of $\boldsymbol{X}$. An obvious approach for testing this hypothesis is to use the Cox score test of $\boldsymbol{\beta} = 0$ in the Cox-model $\lambda(t|\boldsymbol{X}) = \lambda_0(t)\exp(\boldsymbol{\beta X})$. A robust variant of this test is obtained by replacing all observed covariates by their rank, ranking each component of $\boldsymbol{X}$ separately. If we want to test the null hypothesis of no covariate effect against the alternative of a linear model, $\lambda(t|\boldsymbol{X}) = \beta_0(t) + \boldsymbol{\beta}(t)\boldsymbol{X}$, a test of $\boldsymbol{\beta}(t) = 0$ in this linear model (Aalen, 1980, 1989) could be used.

The covariate order test approach can also in principle be generalized to construct tests of covariate effect in the case with several covariates. The covariate order method for exponential regression can be generalized to situations with a, say, p-dimensional covariate $\boldsymbol{X}$, yielding a point process in $\mathcal{R}^p$ (such a generalization is indicated in Kvaløy and Lindqvist, 1998a). If Z is exponentially distributed and there is no association between Z and $\boldsymbol{X}$, this point process will be approximately a HPP in $\mathcal{R}^p$. Then using a test of the null hypothesis of a p-dimensional HPP on the process in $\mathcal{R}^p$, yields a test of the hypothesis of no covariate effect. Generalizations of the test to other life time distributions can be done by resampling. The practical usefulness of this generalization of the covariate order test approach has yet not been exploited.

Rather than testing no covariate effect versus any type of effect, it is often of more interest to test the significance of one covariate after having taken account for the effect of the other covariates. Let X be the covariate of interest, and let $\boldsymbol{X}' = \boldsymbol{X} \backslash X$ be the vector of the other covariates. An obvious approach for testing the significance of X after having taken account for the effect of $\boldsymbol{X}'$, is to fit the Cox-model $\lambda(t|\boldsymbol{X}) = \lambda_0(t)\exp(\beta X + \boldsymbol{\beta}'\boldsymbol{X}')$, and use standard likelihood method to test the null hypothesis $\beta = 0$. An alternative approach is to fit the linear model $\lambda(t|\boldsymbol{X}) = \beta_0(t) + \beta(t)X + \boldsymbol{\beta}'(t)\boldsymbol{X}'$ and test if $\beta(t) = 0$ in this model (Aalen, 1980, 1989; Grønnesby, 1997). A different idea was presented by O'Brien (1978) who suggested first linearly regressing X on $\boldsymbol{X}'$, and then using the residuals of this regression instead of X, and test if these residuals constitute a significant covariate.

O'Quigley and Prentice (1991) suggest an alternative Cox-model approach in which $\boldsymbol{\beta}'\boldsymbol{X}'$ is replaced by $\Psi(\boldsymbol{X}')$ which is a function arrived at by exploratory techniques in order to obtain the best-fitting model including $\boldsymbol{\beta}'\boldsymbol{X}'$ only. Moreover, for making the procedure robust, they replaced the observed covariate X_i by a function $\gamma(X_i)$ of the rank of X_i (logit rank), before testing if $\beta = 0$ in the model $\lambda(t|\boldsymbol{X}_i) = \lambda_0(t)\exp(\beta\gamma(X_i) + \alpha\Psi(\boldsymbol{X}_i'))$. Another test approach suggested by O'Quigley and Prentice (1991) is to stratify on the covariates $\boldsymbol{\beta}'\boldsymbol{X}_i'$, and defining, say, m classes. Then test if $\beta = 0$ in the model $\lambda(t|X_i) = \lambda_{0j}(t)\exp(\beta\gamma(X_i))$, $i = 1, \ldots m$. This approach makes weaker assumptions than the approach based on the Cox-model.

A third test approach based on a Cox-model assumption is as follows. First estimate the model $\lambda(t|\boldsymbol{X}) = \lambda_0(t)\exp(g(X) + g'(\boldsymbol{X}'))$. Here $g(\cdot)$ and $g'(\cdot)$ are either parametric functions or arbitrary smooth nonparametric functions. Next transform the observation times

$T_1, \ldots, T_n$ to $L_1, \ldots, L_n$, where $L_i = \hat{\Lambda}_0(T_i) \exp(\hat{g}'(\boldsymbol{X}'))$ and $\hat{\Lambda}_0(t) = \int_0^t \hat{\lambda}_0(u) du$. Then test if there is a covariate effect in the sample $(L_1, \delta_1, X_1), \ldots, (L_n, \delta_n, X_n)$ for instance by using the bootstrapped AD test. The motivation for this test approach is the observation that if Z has the true hazard rate $\lambda(t|\boldsymbol{X}) = \lambda_0(t) \exp(g(X) + g'(\boldsymbol{X}'))$, then $L = \Lambda_0(Z) \exp(g'(\boldsymbol{X}'))$ will be exponentially distributed with parameter $\exp(g(X))$. Thus a test for covariate effect based on the sample $(L_1, \delta_1, X_1), \ldots, (L_n, \delta_n, X_n)$ will be a test of the significance of X when the effect of $\boldsymbol{X}'$ is accounted for. In practice, the transformed sample $L_1, \ldots, L_n$ will only approximately be a sample of censored exponentially distributed variables. Thus using the tests of Section 2.1 to test the significance of X is not recommended. Tests with the more general null hypothesis of any life time distribution will also be more robust against deviations from the model assumptions. For practical use of this test approach in a nonparametric Cox-model, see Kvaløy (1999).

A test approach similar to the approach above is to instead of estimating the full model, start by estimating the model $\lambda(t|\boldsymbol{X}') = \lambda_0(t) \exp(g'(\boldsymbol{X}'))$ where X is not included. Under the null hypothesis, $L = \Lambda_0(Z) \exp(g'(\boldsymbol{X}'))$ will be standard exponentially distributed. Then letting $L_i = \hat{\Lambda}_0(T_i) \exp(\hat{g}'(\boldsymbol{X}'))$ for $i = 1, \ldots, n$, a test of the significance of X can be based on $(L_1, \delta_1, X_1), \ldots, (L_n, \delta_n, X_n)$. Notice the difference between this approach and the similar approach above. In this case the test is based on the "residuals" $L_1, \ldots, L_n$ of a model not including X, while the test above is test of the significance of X in a model including both X and $\boldsymbol{X}'$.

7 Conclusion

The covariate order approach generates new and interesting tests for covariate effect. One of these tests, the resampled AD test is, based on the simulation study in this paper, the recommended test for general use among the tests studied. The resampled AD test has good power properties both against monotonic and nonmonotonic alternatives and is robust against covariate outliers and misspecifications of the covariate scale.

When a resample test is used, it is in general indifferent if the bootstrap test or the permutation test is used. The only exception is in cases with very few observations, then a permutation test is preferable. All resample tests requires the assumption of random censoring.

Both the simulation study and some of the examples suggest that robust rank-based tests generally are preferable to covariate-based tests. For the rank-based tests the scale on which covariates are measured is of no concern, and a moderate amount of outliers will only have a limited influence on the tests. Moreover, despite the fact that the rank-based tests are not using all available information by only using the rank of the covariate values, there seems in general to be relatively small differences in power between covariate-based and rank-based tests.

Acknowledgements

I would like to thank my advisor Bo Lindqvist for a number of very helpful discussions and suggestions during the work on this paper. Also thanks to Eirik Skogvoll for providing the cardiac arrest data. This work was funded by a PhD grant from the Research Council of Norway.

References

Aalen, O. O. (1980). A model for nonparametric regression analysis of counting processes, *Lecture Notes in Statistics*, Vol. 2, Springer, New York, pp. 1–25.

Aalen, O. O. (1989). A linear regression model for the analysis of life times, *Statistics in Medicine* **8**: 907–925.

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.

Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit, *Journal of the American Statistical Association* **49**: 765–769.

Ascher, H. and Feingold, H. (1984). *Repairable Systems Reliability. Modeling, Inference, Misconceptions and Their Causes*, Marcel Dekker, Inc., New York.

Bain, L. J., Engelhardt, M. and Wright, F. T. (1985). Tests for an increasing trend in the intensity of a Poisson process: A power study, *Journal of the American Statistical Association* **80**: 419–422.

Beran, R. (1986). Simulated power functions, *The Annals of Statistics* **14**: 151–173.

Beran, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements, *Journal of the American Statistical Association* **83**: 687–697.

Brown, Jr., B. W., Hollander, M. and Korwar, R. M. (1974). Nonparametric tests of independence for censored data, with applications to heart transplant studies, *in* F. Proschan and R. J. Serfling (eds), *Reliability and Biometry : Statistical Analysis of Lifelength*, Philadelphia: SIAM, pp. 327–354.

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

Cox, D. R. and Lewis, P. A. W. (1966). *The Statistical Analysis of Series of Events*, Methuen, London.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*, London: Chapman & Hall.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*, Cambridge Series on Statistical and Probabilistic Mathematics, Cambridge University Press.

Elvebakk, G. (1998). Robustification of trend tests by resampling techniques, *in* S. Lydersen, G. K. Hansen and H. A. Sandtorv (eds), *Proceedings of ESREL '98*, Vol. 2, Balkema, Rotterdam, Trondheim, Norway, 16-19 June, pp. 1365–1372.

Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information, *Biometrics* **21**: 826–838.

Grønnesby, J. K. (1997). Testing covariate effects in Aalen's linear hazard model, *Scandinavian Journal of Statistics* **24**: 125–135.

Jones, M. P. (1991). Robust tests for survival data involving a single continuous covariate, *Scandinavian Journal of Statistics* **18**: 323–332.

Jones, M. P. and Crowley, J. (1989). A general class of nonparametric tests for survival analysis, *Biometrics* **45**: 157–170.

JONES, M. P. AND CROWLEY, J. (1990). Asymptotic properties of a general class of nonparametric tests for survival analysis, *The Annals of Statistics* **18**: 1203–1220.

KVALØY, J. T. (1999). Nonparametric estimation in Cox-models: Time transformation methods versus partial likelihood methods, *Technical Report 3*, Norwegian University of Science and Technology, Department of Mathematical Sciences.

KVALØY, J. T. AND LINDQVIST, B. H. (1998a). The covariate order method for censored exponential regression, *Technical Report 10*, Norwegian University of Science and Technology, Department of Mathematical Sciences.

KVALØY, J. T. AND LINDQVIST, B. H. (1998b). TTT-based tests for trend in repairable systems data, *Reliability Engineering and System Safety* **60**: 13–28.

LE, C. T. AND GRAMBSCH, P. M. (1994). Tests of association between survival time and a continuous covariate, *Communications in Statistics - Theory and Methods* **23**: 1009–1019.

LINDQVIST, B., KJØNSTAD, G. A. AND MELAND, N. (1994). Testing for trend in repairable systems data, *Proceedings of ESREL '94*, La Baule, France, 30 May-3 June, pp. 488–496.

MCKEAGUE, I. W., NIKABADZE, A. M. AND SUN, Y. (1995). An omnibus test for independence of a survival time from a covariate, *The Annals of Statistics* **23**: 450–475.

O'BRIEN, P. C. (1978). A nonparametric test for association with censored data, *Biometrics* **34**: 243–250.

O'QUIGLEY, J. AND PRENTICE, R. L. (1991). Nonparametric tests of association between survival time and continuously measured covariates: The logit-rank and associated procedures, *Biometrics* **47**: 117–127.

ROMANO, J. (1989). Bootstrap and randomization tests of some nonparametric hypotheses, *The Annals of Statistics* **17**: 141–150.

SKOGVOLL, E. AND LINDQVIST, B. H. (1999). Modeling the occurrence of cardiac arrest as a Poisson process, *Annals of Emergency Medicine*. In press.