Covariate Order Tests for Covariate Effect

JAN TERJE KVALØY

Abstract. A new approach for constructing tests for 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 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 only use the ordering of the covariate. Of particular interest is a test based on the Anderson-Darling statistic. This test has good power properties both against monotonic and nonmonotonic dependencies between the covariate and the life time variable.

Keywords: Tests for association, rank-based tests, permutation tests, nonmonotonic effects, censored data, point processes

1. Introduction

The basic problem studied in this paper is that of testing if there is an association between the nonnegative random variable Z and the random variable X, where the first variable may be subject to right censoring by the random variable C. For convenience we will use the life time data terminology and call Z the *life time*, C the *censoring time*, $T = \min(Z, C)$ the *observation time* and X the *covariate*. 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.

For $i = 1, \ldots, n$, the censoring time C_i is generally assumed to be conditionally independent of the life time Z_i given the covariate X_i, and to have a cumulative distribution function $F_{C_i}(c|X_i)$ given X_i. This will be called *independent censoring* (see for instance Kalbfleisch and Prentice, 1980). In some discussions the more restrictive assumption that the censoring time C_i is independent of both the life time Z_i and the covariate X_i and has the same cumulative distribution function $F_C(c)$ for all i, is made. This will be called *random censoring*. The life time Z_i is assumed to come from a distribution with finite first order moment, support on the positive real line and cumulative distribution function $F_Z(z|X_i)$ given X_i. In general X_i can be a vector of possibly time dependent covariates, but for simplicity we shall only consider the case of a single covariate which is constant in time. For a discussion of generalizations to the case of several covariates, see Kvaløy (1999).

Tests for covariate effect have been discussed by a number of authors. Jones (1991) studied 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 by O'Quigley and Prentice (1991), the Brown, 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 and Crowley class of test statistics, in particular a modified generalized log-rank test. The tests for covariate effect in the Jones and 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, 1989; 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 have been suggested by Le and Grambsch (1994) and McKeague et al. (1995). These approaches do, however, have some limitations. 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 and generally having low power.

In the present paper a new approach for constructing tests for covariate effect called the *covariate order* approach is proposed. The idea of this approach is arising from the derivation of a method for exponential regression presented by Kvaløy and Lindqvist (1998a). First consider the case with no censoring. The idea is to first arrange the observations $(Z_1, X_1), \ldots, (Z_n, X_n)$ such that $X_1 \leq X_2 \leq \cdots \leq X_n$. Observations with equal covariate values are arranged in random order. Then construct a point process on the line in which the life times $Z_1, \ldots, Z_n$ are subsequent inter-arrival times. It is easily realized that this point process will be a renewal process (RP) under the null hypothesis of no covariate effect since all the inter-arrival times then will be independent and identically distributed. If, on the other hand, there is a covariate effect the constructed process will not be an RP. If for example the life time Z tend to become shorter as the covariate X increases, the inter-arrival times $Z_1, \ldots, Z_n$ will tend to become shorter and shorter. Thus tests of the null hypothesis RP versus the alternative not RP applied to the constructed process will in fact be tests of covariate effect. Tests of RP versus not RP are for instance much studied in the reliability literature, see for example Ascher and Feingold (1984) or Elvebakk (1999) and references therein. Such tests can now, by the suggested construction, be applied as tests for covariate effect.

For the case with censored data, again 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 construct a point process where the observation times $T_1, \ldots, T_n$ are subsequent intervals. In this process, let only points which are endpoints of intervals corresponding to *uncensored* observations be considered as events, occurring at times denoted $S_1, \ldots, S_m$, where $m = \sum_{j=1}^{n} \delta_j$. 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 $k(i) = \min\{r | \sum_{j=1}^{r} \delta_j = i\}$.

In the special case of random censoring, the increments $S_i - S_{i-1}$, $i = 1, \ldots, m$, of the process $S_1, \ldots, S_m$ will still be independent and identically distributed under the null hypothesis of no covariate effect. This follows since in this case all the life times $Z_1, \ldots, Z_n$ will be independent and identically distributed and the same is true for all the censoring times $C_1, \ldots, C_n$. Thus tests of RP versus not RP applied to the process $S_1, \ldots, S_m$ can still be used as tests for covariate effect in the case of random censoring.

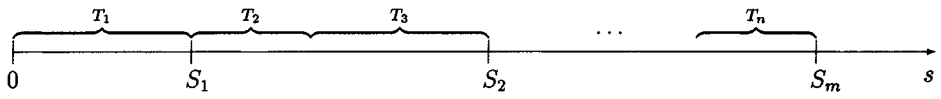

Figure 1. Construction of the process $S_1, \ldots, S_m$.

In the case of independent censoring, however, C may depend on X and this will generally imply that $S_1, \ldots, S_m$ is not an RP even if Z is independent of X. There is, however, one important exception, and this is in the case of exponentially distributed life times. Under the assumption of independent censoring and exponentially distributed life times the process $S_1, \ldots, S_m$ will in fact be a homogeneous Poisson process (HPP) under the null hypothesis of no covariate effect (see Section 2.1). Thus in the case of exponentially distributed life times we can use tests constructed for testing HPP versus not HPP as tests for covariate effect. Tests of HPP versus not HPP are for instance widely studied in the reliability literature, see for example Ascher and Feingold (1984) or Kvaløy and Lindqvist (1998b) and references therein.

An obvious way of extending the covariate order test approach to the assumption of independent censoring is thus to transform the observation times such that the transformed life times becomes exponentially distributed. This is theoretically achieved by transforming the observation times by the integrated hazard rate of the life time distribution. In practice the integrated hazard rate is not known but can be consistently estimated. Thus in the general case, testing for covariate effect is done by constructing the process $S_1,\ldots, S_m$ using the transformed observation times and applying tests of the null hypothesis HPP to this process. In addition some refinements using resampling methods are useful.

Tests derived by the covariate order approach are robust against outliers in the covariate space since they are only using the ordering of the recorded covariate values. Of particular interest is a test based on the Anderson-Darling statistic (Kvaløy and Lindqvist, 1998b). This test has favorable properties as a test for covariate effect since it unlike most other tests for covariate effect has good power properties both against monotonic and non-monotonic alternatives.

Notice that no assumption of continuous X is needed for constructing covariate order tests. Covariate order tests are obviously generally not well suited for taking into account the effect of time dependent covariates. Basing a test on the ordering of the covariates at time 0 may however work reasonably well in certain cases.

The covariate order test approach is presented in Section 2 and a resampling method which can be used to improve the level properties of the test approach is presented in Section 3. A simulation study exploring the small sample properties of a number of tests for covariate effect is presented in Section 4. Some real data examples are presented Section 5 and, finally, some concluding comments are given in Section 6.

2. The Covariate Order Test Approach

In this section the covariate order test approach is presented. First in Section 2.1 the special case of exponentially distributed life times is discussed, and two concrete examples of covariate order tests are presented. The generalization to any life time distribution is discussed in Section 2.2, and some comments are given in Section 2.3.

2.1. Exponentially Distributed Life Times

We now consider the case when Z, conditionally given X, is exponentially distributed with hazard rate $\lambda(X)$. For this case we want to test the null hypothesis of no covariate effect, or in other words the null hypothesis that $\lambda(X) \equiv \lambda$.

Based on the observations $(T_1, \delta_1, X_1),\ldots, (T_n, \delta_n, X_n)$ the process $S_1,\ldots, S_m$ is constructed as explained in Section 1. Under the null hypothesis of no covariate effect and the assumption of independent censoring this process will be an HPP. A formal proof of this claim is given in the Appendix. An easy informal argumentation is the following: Intuitively the conditional intensity of the process $S_1,\ldots, S_m$ at any point s, given the history of the process until that point, will equal λ. This implies that the process $S_1,\ldots, S_m$ will be an HPP. If, on the other hand there is a covariate effect, which means that $\lambda(x)$ is not constant in x, then the process $S_1,\ldots, S_m$ will not be an HPP.

Kvaløy and Lindqvist (1998a) argued that the process $S_1, \ldots, S_m$ 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. There exist a number of such tests in the literature, see for instance Ascher and Feingold (1984) or Kvaløy and Lindqvist (1998b) and references therein. Two such tests are presented below.

One of the most popular and frequently used tests in the HPP versus NHPP setting is the Laplace test. Let $S = \Sigma_{i=1}^{n} T_i$ and for convenience define

$$\hat{m} = \begin{cases} m & \text{if } S_m < S \\ m-1 & \text{if } S_m = S \end{cases}$$

Then the test statistic of the Laplace test is

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

This statistic is approximately normally distributed under the null hypothesis that $S_1, \ldots, S_m$ is an HPP, and the approximation is very good even for very small samples. The Laplace test has optimal properties against certain monotonic alternatives, see for instance Bain et al. (1985).

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 conditional intensity of the process $S_1, \ldots, S_m$ will be monotonic, 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 $S_1, \ldots, S_m$ is nonmonotonic.

Kvaløy and Lindqvist (1998b) studied a test of HPP versus NHPP based on the Anderson-Darling statistic. 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

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

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 by Anderson and Darling (1954). The asymptotic distribution is a good approximation to the exact distribution even for sample sizes smaller than 10. On a 5% level the null hypothesis is rejected if $AD \geq AD_{0.05} = 2.492$.

2.2. General Life Time Distributions

In the general case, when Z given X can have any life time distribution, the basic idea is to transform the observation times to a sample of censored approximately exponentially distributed life times and then apply the tests presented in Section 2.1 to the transformed sample. This approach is outlined below.

Assume that the null hypothesis of no covariate effect holds. Let $\lambda(t)$ be the hazard rate of the distribution of Z and define the integrated hazard rate $\Lambda(t) = \int_0^t \lambda(u)du$. Then the transformed variable $\Lambda(Z)$ will be standard exponentially distributed. Thus for known $\Lambda(t)$, transforming the observation times to $\Lambda(T_1),\ldots,\Lambda(T_n)$ would yield a censored sample from the standard exponential distribution, and any test for covariate effect constructed for exponentially distributed data could be applied to this sample. In practice $\Lambda(t)$ is unknown but can under the null hypothesis be consistently estimated by the Nelson-Aalen estimator, $\hat{\Lambda}(t) = \Sigma_{i=1}^n \delta_i [\Sigma_{j=1}^n I(T_j \geq T_i)]^{-1} I(T_i \leq t)$. See for instance Andersen et al. (1993) for details.

Thus a reasonable test procedure will be to apply tests for covariate effect constructed for exponentially distributed data to the transformed observations $(\hat{\Lambda}(T_1), \delta_1, X_1),\ldots,(\hat{\Lambda}(T_n), \delta_n, X_n)$. A drawback with this approach is the loss of information introduced by replacing the continuous observation times $T_1,\ldots,T_n$ by the discrete transformation $\hat{\Lambda}(T_1),\ldots,\hat{\Lambda}(T_n)$. On the other hand, this discretization also implies robustness against outliers in the recorded observation times. Another drawback is that the transformation implies certain dependencies in the transformed observations. These effects diminishes with increasing sample size.

Despite the certain loss of information, dependencies, and the fact that the transformed life times are only approximately exponentially distributed, this test approach works fairly well in practice (see Section 4). However, in particular for small sample sizes, the resampling method presented in Section 3 could and should be used to improve the level properties of the tests. This method does in particular resolve the dependency problem.

2.3. Comments

Replacing the observation times by the discrete transformation $\hat{\Lambda}(T_1),\ldots,\hat{\Lambda}(T_n)$ may be viewed as a natural extension of the exponential ordered scores proposed by Cox (1964). Cox (1964) proposed to replace observations which are assumed to be exponentially distributed by their exponential scores to obtain robustness against deviations from the assumption of exponentiality. The exponential ordered scores are calculated by ranking the observations and replacing each observation by the expected value of the corresponding order statistics of the standard exponential distribution. If there are no censored observations the transformation $\hat{\Lambda}(T_1),\ldots,\hat{\Lambda}(T_n)$ corresponds exactly to the exponential ordered scores of $T_1,\ldots,T_n$, while the case with censored observations corresponds to a natural extension of the exponential ordered scores (see for instance Nelson, 1972).

3. Resampling

For small or moderate sample sizes it is recommended to use a resampling version of the test approach presented in Section 2.2. This yields tests with improved level properties. Both bootstrap and permutation methods can be used, below a permutation method is discussed.

The idea is the following. First assume that $\Lambda(t)$ is known and consider the transformed sample $(\Lambda(T_1), \delta_1, X_1),\ldots,(\Lambda(T_n), \delta_n, X_n)$. By the arguments of Section 2.1 this transformed sample will yield a process $S_1,\ldots,S_m$ which under the null hypothesis is an HPP. Further, under the null hypothesis any permutation of the transformed sample, $(\Lambda(T_1), \delta_1, X_{\pi(1)}),\ldots,(\Lambda(T_n), \delta_n, X_{\pi(n)})$, will also yield a process $S_1^\star,\ldots,S_m^\star$ which is an HPP and any such permutation is equally likely. Here $\pi(1),\ldots,\pi(n)$ denotes a permutation of the numbers $1,\ldots,n$. With $\Lambda(t)$ replaced by $\hat{\Lambda}(t)$, any permutation of the transformed observations will yield a process which under the null hypothesis is approximately an HPP.

Let the test statistic of a test for covariate effect constructed for exponentially distributed life times be denoted TE. For simplicity we assume that the the null hypothesis is rejected for large values of TE (if for example the Laplace test is used let $TE = |LAP|$). Let TE_{obs} be the observed value of TE calculated from the transformed observations $(\hat{\Lambda}(T_1), \delta_1, X_1), \ldots,$ $(\hat{\Lambda}(T_n), \delta_n, X_n)$ and let $TE^\star$ be the value of TE calculated from a permutation of the transformed observations, $(\hat{\Lambda}(T_1), \delta_1, X_{\pi(1)}), \ldots, (\hat{\Lambda}(T_n), \delta_n, X_{\pi(n)})$. The exact permutation null distribution of TE is found by calculating $TE^\star$ for all $n!$ possible permutations of the original observations. In practice it is sufficient to calculate $TE^{\star(1)}, \ldots, TE^{\star(P)}$ for a large number, P, of randomly selected permutations, and calculate the approximate p-value $\hat{p} = \sum_{i=1}^{P} I(TE^{\star(i)} > TE_{obs}) / P$. For simulations choosing P equal to 1000 (or less) is in most cases sufficient (see Davison and Hinkley, 1997, Chapter 4.2.5). For calculating p-values for real data somewhat larger values are recommended if the p-value is small.

Notice that the order of permutation and transformation is indifferent. We might think of the permutation test as first permuting the original observations and then transforming the observation times before calculating the test statistic. The order is indifferent since the estimated integrated hazard rate of course will be the same for any permutation of the observations.

In the special case of no or random censoring, the permutation test approach can be applied directly to the original observations $(T_1, \delta_1, X_1), \ldots, (T_n, \delta_n, X_n)$ without doing any transformation since there in this case is no potential covariate effect in the censoring variable which needs to be taken care of. Any permutation $(T_1, \delta_1, X_{\pi(1)}), \ldots, (T_n, \delta_n, X_{\pi(n)})$ is equally likely under the null hypothesis, and the permutation test is in this case an exact conditional combinatorial test for covariate effect. See for instance Romano (1989) or Davison and Hinkley (1997) for detailed discussions on permutation tests.

In the general case the observation times $T_1, \ldots, T_n$ needs to be replaced by the transformed observation times to cope with the possible covariate effect in the censoring variable. This possible covariate effect is then masked by the process $S_1, \ldots, S_m$ constructed from the transformed observation times being approximately an HPP even if there is an covariate effect in the censoring variable. The permutation test based on the transformed observations is an exact conditional combinatorial test, but generally strictly speaking not purely a test of covariate effect in the life time variable. Since the transformation $\hat{\Lambda}(t)$ yields transformed life times which are only approximately exponentially distributed the effect of a potential covariate effect in the censoring time is not necessarily completely masked, and the test might have a certain remaining sensitivity to covariate effects in the censoring variable. In practice, however, it turns out that there is no such sensitivity, the transformation $\hat{\Lambda}(t)$ successfully masks even very strong covariate effects in the censoring distribution. Finally notice that since the permutation test is conditional on the observations and the transformed observation times are the same for any permutation of the original observations, the dependencies introduces in these transformed observation times cause no concern in the permutation test.

Resampling methods similar to the method discussed above do of course also apply to other tests for covariate effect than covariate order tests, and may for instance be used for improving the level properties of tests with critical values based on the asymptotic distribution. In the case of no or random censoring the permutation method apply to any test of covariate effect. The method also often apply under the assumption of independent censoring by similar arguments as used for the covariate order tests.

4. Simulation Study

In this section the two covariate order tests suggested in Section 2 are compared to some of the tests mentioned in Section 1 in a simulation study. The other tests considered are the two tests recommended for general use by Jones (1991), the Cox score test (Cox, 1972) and the modified generalized log-rank test (Jones, 1991), and a standard test constructed to have power both against monotonic and nonmonotonic covariate effects. This test is based on dividing the covariate axis into q intervals and introducing the q indicator variables $I_1, \ldots, I_q$ where $I_i(x) = 1$ if x is in the ith interval, and 0 otherwise. Then fitting the Cox-model

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

and using the Cox score test to test the null hypothesis that $\beta_1 = \ldots = \beta_{q-1} = 0$, 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. Preliminary simulations indicated that the best approach for the cases studied in this simulation study is to divide the covariate axis into three intervals such that 1/3 of the observations fall in each interval.

The abbreviations COX for the Cox score test, MGL for the modified generalized log-rank test and COX3 for the test based on dividing the covariate axis into three intervals are introduced. Critical values for these tests are based on the asymptotic distributions. For the covariate order tests presented in Section 2, the abbreviations AD for the Anderson-Darling test for trend (2) and LAP for the Laplace test (1) applied to transformed observations $(\hat{\Lambda}(T_1), \delta_1, X_1), \ldots, (\hat{\Lambda}(T_n), \delta_1, X_n)$ are introduced.

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 permutation tests 1000 permutations are generated. 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 permutation versions of all tests are also 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.

We see that not all of the ordinary tests achieve the correct level for small samples. The AD and LAP tests are in fact too conservative, while the COX3 test and to some extent the COX test, are nonconservative. For increasing sample sizes the level properties of the tests improve. For $n = 50$ the deviations from the correct level are not large, but the AD test and the LAP tests are still slightly conservative while the COX3 test is slightly non-conservative. For all tests, the permutation version of the test achieves the correct level in all cases. Subsequently only the permutation versions of the covariate order tests will be studied, and these will be denoted respectively AD-perm and LAP-perm.

Another illustration of level properties is presented in Figure 2. In this example samples of size 50 are generated using a life time distribution with hazard rate $\lambda(t|x) = 2at$ and a censoring distribution with hazard rate $\gamma(t|x) = 2\exp(bx)t$. In other words a situation with no covariate effect in the life time distribution, but dependence between the censoring variable and the covariate. The covariates are drawn from a uniform[0,1] distribution and a is adjusted according to b to give approximately 50% censoring in all cases.

Table 1. Simulated rejection probabilities 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 and permutation tests are reported.

	$n = 10$		$n = 30$		$n = 50$	
Test	Ord.	Perm.	Ord.	Perm.	Ord.	Perm.
AD	0.012	0.054	0.036	0.053	0.040	0.049
LAP	0.022	0.053	0.042	0.053	0.041	0.047
COX	0.062	0.051	0.055	0.051	0.056	0.051
MGL	0.049	0.051	0.052	0.050	0.051	0.047
COX3	0.082	0.053	0.067	0.048	0.063	0.054

Figure 2 illustrates that the AD-perm and LAP-perm tests remain on the 5% level for all values of b, in other words even in cases with very strong covariate effects in the censoring distribution. The COX and MGL tests also achieve the correct level in all cases, while the COX3 test is slightly non-conservative.

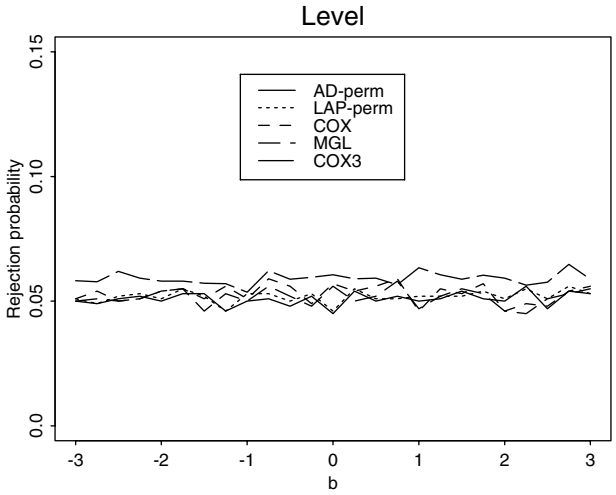

Figure 2. Simulated rejection probabilities as a function of b for samples of 50 observations from a model with hazard rate $2at$ for the life time variable and hazard rate $2\exp(bx)t$ for the censoring variable, and with a adjusted to give 50% censoring. Lines are drawn between the estimates of the rejection probability for different values of b.

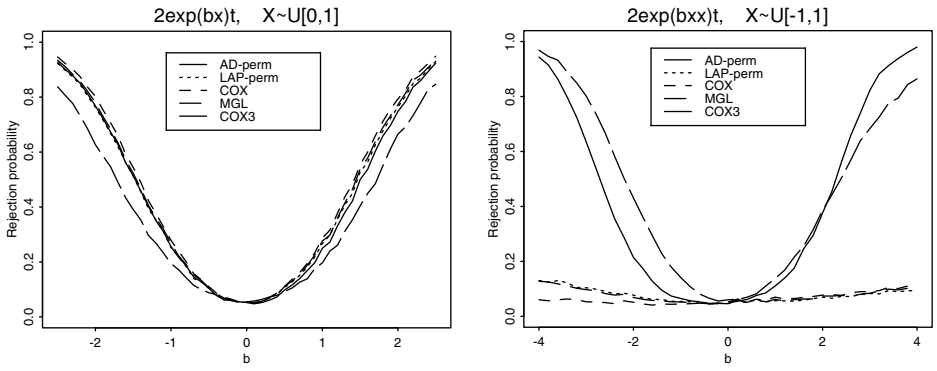

Figure 3. Estimated rejection probabilities as a function of b for samples of 50 observations from models with hazard rates $2\exp(bx)t$ and $2\exp(bx^2)t$, covariates drawn respectively from a uniform[0,1] and a uniform[−1,1] distribution. The hazard rates of the censoring variables are respectively $2\exp(a - bx)t$ and $2\exp(a - bx^2)t$, and for each value of b, a is chosen to give approximately 50% censoring. Lines are drawn between estimates of the rejection probability for different values of b.

4.2. Power

Figure 3 illustrates power properties of the various tests for models with hazard rate respectively $\lambda(t|x) = 2\exp(bx)t$ and $\lambda(t|x) = 2\exp(bx^2)t$. The covariates are drawn respectively from a uniform[0,1] and a uniform[−1,1] distribution. The corresponding censoring variables are drawn from densities with hazard rates respectively $\gamma(t|x) = 2\exp(a - bx)t$ and $\gamma(t|x) = 2\exp(a - bx^2)t$ where a is chosen for each value of b to give approximately 50% censoring. Samples of size 50 are generated.

The left plot 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.

The right plot in Figure 3 illustrates the power properties of the tests in the 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 the AD-perm and the COX3 test have good power properties against this nonmonotonic alternative as well.

The two plots in Figure 3 illustrate that the tests are able to detect the covariate effect in the life time distribution even if the covariate effect in the censoring distribution is in the "opposite direction."

Another example of 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 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 AD-perm test and the COX3 test are far more powerful than the other tests.

Table 2. Simulated rejection probabilities 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.

Test	U[0,0.25]	U[0,0.5]	U[0,0.75]	U[0,1]
AD-perm	0.492	0.990	0.917	0.688
LAP-perm	0.509	0.989	0.705	0.021
COX	0.521	0.995	0.850	0.039
MGL	0.495	0.991	0.795	0.040
COX3	0.395	0.974	0.904	0.829

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 $\lambda(t|x)$. This is a situation for which the COX3 test is particularly well suited. The test is slightly modified by instead of dividing the observations into three groups such that each group contains 1/3 of the observations, the observations are divided into three groups according to the three covariate values.

The estimated rejection probabilities are reported in Table 3.

The first column of this table shows that the covariate order tests still achieve the correct level, while the MGL test is slightly too conservative and the COX and, in particular, the COX3 tests are slightly non-conservative. Otherwise the table shows that rank-based tests (all tests except the COX test are rank-based), including the covariate order tests, works well even in cases with discrete covariates. In the cases with nonmonotonic covariate effects, the AD-perm test and the COX3 test are far more powerful than the other 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) studied the

Table 3. Simulated rejection probabilities 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}.

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-perm	0.049	0.788	0.788	0.670	0.971	0.673	0.798
LAP-perm	0.048	0.790	0.020	0.664	0.021	0.679	0.018
COX	0.058	0.853	0.024	0.761	0.035	0.768	0.025
MGL	0.038	0.793	0.026	0.691	0.036	0.691	0.027
COX3	0.073	0.775	0.999	0.688	1.000	0.673	0.999

Table 4. Simulated rejection probabilities 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.

Covar. dist.	Type of contam.	AD-perm	LAP-perm	COX	MGL	COX3
exp(2)	1	0.757	0.748	0.852	0.782	0.651
	2	0.582	0.580	0.311	0.594	0.485
	3	0.639	0.644	0.434	0.648	0.528
	4	0.775	0.769	0.797	0.793	0.664
N(2,1)	1	0.761	0.759	0.772	0.759	0.664
	2	0.614	0.599	0.531	0.606	0.512
	3	0.617	0.594	0.363	0.591	0.510
	4	0.752	0.749	0.638	0.708	0.663
U[0,2]	1	0.843	0.849	0.864	0.841	0.704
	2	0.669	0.649	0.401	0.652	0.547
	3	0.664	0.647	0.355	0.640	0.546
	4	0.837	0.853	0.824	0.835	0.710

robustness of tests for covariate effect under different models for covariate contamination. The simulation study presented by 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 4.

As expected, we see from Table 4 that the COX test, which uses the actual reported values of the covariates, generally is less robust against covariate contamination than the other rank-based tests. The covariate order tests and the MGL test have fairly equal and generally good power properties. The COX3 test is in all cases considered in Table 4 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

Table 5. Test statistics and p-values for the glioma data. 10000 repetitions are used for estimating the p-values of the permutation tests.

	Unmodified data		Modified data	
Test	Statistic	p-value	Statistic	p-value
AD-perm	3.87	0.0054	4.21	0.0036
LAP-perm	2.72	0.0013	2.74	0.0012
COX	3.15	0.0016	1.40	0.1615
MGL	3.16	0.0016	2.72	0.0065
COX3-perm	8.54	0.0201	8.54	0.0203

Anderson-Darling test and the Laplace test have been considered in this simulation study. Depending on which assumptions can be made, different variants of these tests can be used. We have focused on the most general case with independent censoring and general life time distributions in which case the tests should be based on transformed observation times. We have further mainly considered the permutation version of the tests to ensure that we have tests with good level properties for any sample size. Notice, however, that for reasonable sample sizes applying the ordinary AD and LAP tests yield tests which may be slightly too conservative but which for practical purposes are equivalent to the permutation tests. Also recall that in cases with no or random censoring there is no need to transform the data. In such cases the permutation versions of the covariate order tests can be applied directly to the original data, or tests of the null hypothesis RP can be used.

The conclusion of the comparison of the AD-perm and LAP-perm tests to other tests is that the two tests, both with regard to power properties and robustness, are useful alternatives to existing tests. In particular is this the case for the AD-perm test. In terms of power against general alternatives, correct level and robustness, this test seems to be the best test for general use among the tests studied. If we only want or need power against monotonic alternatives, the AD-perm test is still a safe choice, but the LAP-perm test or the MGL 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 COX test generally has at least as good power properties against monotonic covariate effects as the other tests.

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 cannot really be directly 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 has correct level. For this kind of test there is also the problem of how to divide the covariate axis into intervals.

It is demonstrated in the simulation study that covariate order tests and other rank-based tests can be successfully used in situations with discrete covariates. Notice, however, 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. Rather than basing the test

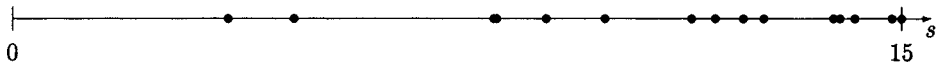

Figure 4. The constructed process $S_1, \ldots, S_m$ for the glioma data.

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 can then be approximated by resampling by taking the same mean or median for all resamplings of the original data.

5. Examples

5.1. Glioma Data

Jones and Crowley (1989) gave 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). For illustrating the effect of extreme covariate values on the various test statistics, Jones and Crowley (1989) considered both the original data set and a modified data set where the age of one of the patients was changed from 57.8 to 97.8. The test statistics and p-values of the various tests, both for the unmodified and the modified data, are reported in Table 5. Since we only have 28 observations, the permutation version of the COX3 test is used to assure that the test will have correct level.

For the unmodified data all tests finds the covariate effect of age to be significant on a 5% level. For the modified data, however, the COX test does not yield a significant result. Also notice that the COX3 test yields clearly larger p-values than the other rank-based tests.

The constructed process $S_1, \ldots, S_m$ based on transformed observation times for the unmodified data is displayed in Figure 4. This figure clearly indicates that the survival times decrease with increasing age at treatment.

Table 6. Test statistics and p-values for the covariate effect of wbc and log(wbc) in the leukemia data. 10000 repetitions are used for estimating the p-values of the permutation tests.

	wbc		log(wbc)	
Test	Statistic	p-value	Statistic	p-value
AD-perm	4.18	0.0029	4.18	0.0036
LAP-perm	2.75	0.0011	2.75	0.0012
COX	2.10	0.0357	3.10	0.0019
MGL	2.76	0.0058	2.86	0.0042
COX3-perm	8.58	0.0278	8.58	0.0247

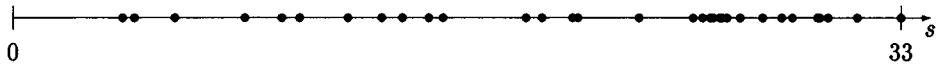

Figure 5. The constructed process $S_1, \ldots, S_m$ for the leukemia data.

This example illustrates the considerable influence of a single extreme covariate value on a test like the COX which uses the actual recorded covariate values, whereas the rank-based tests remain robust.

5.2. Leukemia Data

Feigl and Zelen (1965) presented uncensored data on survival times for patients with leukemia. We here only consider the effect of the covariate white blood cell count (wbc) for all 33 patients. We consider this covariate measured on two different scales, the original measurements of wbc as reported by Feigl and Zelen (1965) and the logarithm of wbc. The results of using the various tests to test the significance of the covariate taken on these two scales are reported in Table 6.

We see that the COX test yields a clearly lower p-value for log(wbc) than for wbc, while the other tests yield the same result for wbc and log(wbc) (the p-values of the permutation tests are estimated separately in each case and thus differ slightly). Notice that the test statistic of the MGL test is slightly different in the two cases. The reason is that this test is not purely rank-based. As in the previous example, the COX3 test does also in this case yield clearly larger p-values than the other tests. The constructed process $S_1, \ldots, S_m$ based on transformed observation times for the leukemia data is plotted in Figure 5. This plot indicates that the survival times decrease with increasing value of wbc.

This example illustrates that rank-based tests are not sensitive to the choice of scale for measuring the covariate, a choice which is not always obvious.

6. Conclusion

The covariate order approach generates new and interesting tests for covariate effect. Covariate order tests are purely rank-based tests and will thus be robust against covariate outliers and misspecifications of the covariate scale as demonstrated in the simulation study and the examples. The AD-perm test is in particular recommended. In addition to good robustness properties, this test has good power properties both against monotonic and nonmonotonic alternatives.

Acknowledgments

I would like to thank Bo Lindqvist for a number of very helpful discussions and suggestions during the work on this paper, and the referees for useful comments which improved the paper. Part of this work was funded by a PhD grant from the Research Council of Norway.

Appendix

Let Z, conditionally given X, be exponentially distributed with hazard rate $\lambda(X)$, and let the process $S_1, \ldots, S_m$ be constructed as described in Section 1. We shall prove that this process under the null hypothesis of no covariate effect, $\lambda(X) \equiv \lambda$, is an HPP.

Let $\mathcal{F}_s$ be the history of the process $S_1, \ldots, S_m$ in the interval $[0,s)$. This history is formally defined as the sub-σ-algebra $\mathcal{F}_s = \sigma\{S_1, \ldots, S_j : S_j \leq s\}$ for $s \geq 0$. Let $N(s) = \Sigma_{i=1}^{n} I(S_i \leq s)$ be the counting process counting events in the process $S_1, \ldots, S_m$, and let $\rho(s|\mathcal{F}_s)$ be the conditional intensity of the process at the point s (see for example Andersen, Borgan, Gill and Keiding, 1993, Page 75). Further consider the process $S_1^*, \ldots, S_n^*$, where $S_j^* = \Sigma_{i=1}^{j} T_i$. Let $N^*(s) = \Sigma_{i=1}^{n} I(S_i^* \leq s)$ be the counting process counting events in this process and define the history $\mathcal{F}_s^* = \sigma\{X_1, \ldots, X_n; (T_j, \delta_j) : \Sigma_{i=1}^{j} T_i \leq s\}$. Note that $X_1, \ldots, X_n$ is contained in all the $\mathcal{F}_s^*$. Clearly $\mathcal{F}_s \subseteq \mathcal{F}_s^*$. The intensity of the process $S_1, \ldots, S_m$ conditional on the history $\mathcal{F}_s^*$ is

$$
\begin{aligned}
&\rho(s|\mathcal{F}_s^*) \\
&\quad = \lim_{\Delta s \to 0} \frac{P(N(s+\Delta s) - N(s) \geq 1 | \mathcal{F}_s^*)}{\Delta s} \\
&\quad = \lim_{\Delta s \to 0} P(s - S_{N^*(s)}^* \leq Z_{N^*(s)+1}^* < s + \Delta s - S_{N^*(s)}^* \cap \\
&\qquad C_{N^*(s)+1} > s + \Delta s - S_{N^*(s)}^* | \mathcal{F}_s^*) \Delta s \\
&\quad = \lim_{\Delta s \to 0} \frac{P(s - S_{N^*(s)}^* \leq Z_{N^*(s)+1}^* < s + \Delta s - S_{N^*(s)}^*)}{\Delta s P(Z_{N^*(s)+1} > s - S_{N^*(s)}^*)} \frac{P(C_{N^*(s)+1} > s + \Delta s - S_{N^*(s)}^*)}{P(C_{N^*(s)+1} > s - S_{N^*(s)}^*)} \\
&\quad = \lambda(X_{N^*(s)+1}) \\
&\quad = \lambda
\end{aligned}
$$

which is intuitive. Since $\mathcal{F}_s \subseteq \mathcal{F}_s^*$ it follows from the innovation theorem (e.g., Andersen et al., 1993, Page 80), that

$$\rho(s \mid \mathcal{F}_s) = \mathrm{E}[\rho(s \mid \mathcal{F}_s^*) \mid \mathcal{F}_s] = \lambda$$

which means that the process $S_1, \ldots, S_m$ is an HPP under the null hypothesis of no covariate effect. We also realize that if there is a covariate effect, implying that $\lambda(x)$ is not constant in x, then the process $S_1, \ldots, S_m$ is not an HPP.

References

O. O. Aalen, "A model for nonparametric regression analysis of counting processes," in *Lecture Notes in Statistics*, W. Klonecki, A. Kozek, and J. Rosiński (eds.), vol. 2 pp. 1–25, Springer-Verlag: New York, 1980.

O. O. Aalen, "A linear regression model for the analysis of life times," *Statistics in Medicine* vol. 8 pp. 907–925, 1989.

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

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

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

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

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

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

D. R. Cox, "Some applications of exponential ordered scores," *Journal of the Royal Statistical Society, Series B* vol. 26 pp. 103–110, 1964.

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

A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Application*, Cambridge University Press: Cambridge, 1997.

G. Elvebakk, "Analysis of Repairable systems data: Statistical inference for a class of models involving renewals, heterogeneity and time trends," Ph.D. Dissertation, Norwegian University of Science and Technology, Trondheim, Norway, 1999.

P. Feigl and M. Zelen, "Estimation of exponential survival probabilities with concomitant information," *Biometrics* vol. 21 pp. 826–838, 1965.

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

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

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

J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data*, Wiley: New York, 1980.

J. T. Kvaløy, "Statistical methods for detecting and modeling general patterns and relationships in lifetime data," Ph.D. Dissertation, Norwegian University of Science and Technology, Trondheim, Norway, 1999.

J. T. Kvaløy and B. H. Lindqvist, "The covariate order method for censored exponential regression," *Tech. Rep #10*, Norwegian University of Science and Technology, Department of Mathematical Sciences, 1998a.

J. T. Kvaløy and B. H. Lindqvist, "TTT-based tests for trend in repairable systems data," *Reliability Engineering and System Safety* vol. 60 pp. 13–28, 1998b.

C. T. Le and P. M. Grambsch, "Tests of association between survival time and a continuous covariate," *Communications in Statistics - Theory and Methods* vol. 23 pp. 1009–1019, 1994.

I. W. McKeague, A. M. Nikabadze, and Y. Sun, "An omnibus test for independence of a survival time from a covariate," *The Annals of Statistics* vol. 23 pp. 450–475, 1995.

W. Nelson, "Theory and applications of hazard plotting for censored failure data," *Technometrics* vol. 14 pp. 945–966, 1972.

P. C. O'Brien, "A nonparametric test for association with censored data," *Biometrics* vol. 34 pp. 243–250, 1978.

J. O'Quigley and R. L. Prentice, "Nonparametric tests of association between survival time and continuously Measured Covariates: The logit-rank and associated procedures," *Biometrics* vol. 47 pp. 117–127, 1991.

J. Romano, "Bootstrap and randomization tests of some nonparametric hypotheses," *The Annals of Statistics* vol. 17 pp. 141–150, 1989.