Applications of Mathematics

Aneta Andrášiková; Eva Fišerová

Behaviour of higher-order approximations of the tests in the single parameter Cox proportional hazards model

Applications of Mathematics, Vol. 65 (2020), No. 3, 229–244

Persistent URL: http://dml.cz/dmlcz/148140

BEHAVIOUR OF HIGHER-ORDER APPROXIMATIONS OF THE TESTS IN THE SINGLE PARAMETER COX PROPORTIONAL HAZARDS MODEL

Aneta Andrášiková, Eva Fišerová, Olomouc

Received December 5, 2019. Published online May 21, 2020.

Abstract. Survival analysis is applied in a wide range of sectors (medicine, economy, etc.), and its main idea is based on evaluating the time until the occurrence of an event of interest. The effect of some particular covariates on survival time is usually described by the Cox proportional hazards model and the statistical significance of the impact of covariates is verified by the likelihood ratio test, the Wald test, or the score test. In addition to standard tests, appropriate higher-order approximations based on Barndorff-Nielsen and Lugannani-Rice formulas are used for more accurate approximations. In this paper, comparison of these tests' size and power for small sample sizes is performed on simulated datasets with various proportions of right-censored data, distributions of baseline hazard functions and different types of covariate—continuous or discrete.

Keywords: survival analysis; likelihood ratio test; wald test; score test; statistical power; adjusted power; higher-order approximation; confidence band

MSC 2020: 62N01, 62N03

1. Introduction

Survival analysis plays an important role in the statistical analysis of time-to-event data sets mainly in medicine [20]; however, applications can also be found, e.g. in economy [18], or materials engineering [14]. Its main idea is based on the survival time, which is the time to the occurrence of an event of interest [11]. Incomplete information about survival time leads to censored observations [17], namely, the most common right-censoring is considered in what follows.

The research has been supported by the project Support of Academic mobility at Palacký University Olomouc (CZ.02.2.69/0.0/0.0/16_027/0008482) and the projects IGA_PrF_2019_006 and IGA_PrF_2020_015.

The effect of some covariates on survival time is usually investigated by the Cox proportional hazards model [11], and its significance is tested by the likelihood ratio, the Wald test and the score test [9]. These tests are asymptotically equivalent; nevertheless, numerically they give different results in applications depending on the data and the sample size [4]. Chandra, Joshi (1983) [10] proved that the score test is more powerful than the other two for large samples. Yi, Wang (2011) [29] compared these tests under a specific design and based on the simulation study, the Wald test was recommended. The mentioned tests represent approximations of the first order. Second-order asymptotic methods [6], [15], [31] provide useful tools for the improvement of the first-order methods. Pierce, Bellio (2015) [22] applied the Barndorff-Nielsen adjustment [3] and the Skovgaard statistic [26], [27] using the nuisance parameters approach and showed that the higher-order improvement is insensitive to the censoring model. Bělašková, Fišerová (2017) [4] showed the improvement of the accuracy of the p-value for small sample sizes using the Barndorff-Nielsen approximation and the Lugannani-Rice approximation [21] in combination with the Wald statistic in the Cox model with one covariate.

The main aim of this paper is to investigate the behaviour of the power and size of tests mentioned above for a single covariate case, extending [4]. Tests comparison is based on an extensive simulation study. There are many procedures on generating survival data, differing mainly in the type and distribution of censoring, the generation of survival times, or the time dependency of the covariates. The basic procedure for the simulation of survival times dealing with fixed covariates and with an exponential, Weibull and Gompertz distribution of survival times is proposed in [5], [23]. Another issue of simulating survival data is related to censoring, e.g. see [1], [12], [19]. The approach described in [28], which guarantees a predetermined probability of censoring for an individual observation, has been applied here.

The paper is organized as follows. Section 2 contains the description of the Cox proportional hazards model together with test statistics and their corresponding second-order approximations. Section 3 deals with the power and size of tests and discusses possible ways of comparison of the studied tests. Section 4 outlines the settings and the results of the performed simulation study. Section 5 is dedicated to conclusions and discussion notes.

2. Test statistics in the Cox proportional hazards model

The Cox proportional hazards model has its base in the hazard function defined [11] as $h(t) = \lim_{\Delta t \to 0} P(t < T \leqslant t + \Delta t \mid T > t)/\Delta t$, where T denotes the time to an event of interest. This function determines the intensity of the occurrence of

the observed event in time, under the condition that the subject survived the time t. Considering one covariate x, the hazard function for the observation i can be defined [11] in the form $h(t, x_i) = h_0(t)\exp(x_i\beta)$, where $h_0(t)$ is an arbitrary baseline hazard function and β is a regression coefficient. The coefficient β is estimated using the maximum likelihood method based on the partial likelihood function [11]

$$L(\beta) = \prod_{i=1}^{n} \Big(\frac{\exp(x_i\beta)}{\sum_{j\in R_i} \exp(x_j\beta)} \Big)^{\delta_i}, \tag{2.1}$$

where n is the number of observations, $R_i = \{j\colon\ t_j \geqslant t_i\}$ is the risk set and δ_i is an indicator function for censoring, i.e., it equals zero for censored observations and one otherwise.

The key assumption of the Cox model is that the hazard ratio is constant over time [24]. In addition to that, the assumption of no ties for the event times, i.e., events are not grouped, is considered. The violation of that assumption could be dealt with using the Breslow approximation [7] in the case of a small proportion of tied data, otherwise using the Efron method, or the exact method [13]. Moreover, independence of the censoring and the survival times is assumed [11].

The null hypothesis $\mathrm{H}_0\colon\ \beta = \beta_0$ against the alternative $\mathrm{H}_\mathrm{A}\colon\ \beta \neq \beta_0$ is tested by the likelihood ratio (Z_LR), the Wald (Z_W), or the score (Z_S) test statistics [9]:

$$\mathrm{Z}_\mathrm{LR} = 2[l(\hat{\beta}) - l(\beta_0)], \tag{2.2}$$

$$\mathrm{Z}_\mathrm{W} = (\hat{\beta} - \beta_0)^2 \cdot \mathcal{J}(\hat{\beta}), \tag{2.3}$$

$$\mathrm{Z}_\mathrm{S} = [l'(\beta_0)]^2 \cdot [\mathcal{J}(\beta_0)]^{-1}, \tag{2.4}$$

where $\hat{\beta}$ denotes the maximum partial likelihood estimator of β, $l(\beta_0)$ stands for the logarithm of the partial likelihood function (2.1) evaluated at the point β_0, $l'(\beta_0)$ stands for the first derivative of the logarithm of (2.1) and $\mathcal{J}(\hat{\beta})$ is the observed Fisher information, given as the second derivative of (2.1) multiplied by minus one. All three test statistics (2.2)–(2.4) have asymptotically a chi-squared distribution with one degree of freedom under H_0. They can also be expressed in the form having asymptotically a standard normal distribution $\mathcal{N}(0, 1)$ under H_0, namely

$$\mathrm{Z}^*_\mathrm{LR} = \mathrm{sgn}(\hat{\beta} - \beta_0)\sqrt{\mathrm{z}_\mathrm{LR}}, \tag{2.5}$$

$$\mathrm{Z}^*_\mathrm{W} = (\hat{\beta} - \beta_0) \cdot [\mathcal{J}(\hat{\beta})]^{1/2}, \tag{2.6}$$

$$\mathrm{Z}^*_\mathrm{S} = l'(\beta_0) \cdot [\mathcal{J}(\beta_0)]^{-1/2}. \tag{2.7}$$

In addition to these tests, it is possible to use higher-order approximations [3], [4], [15], [31] to the cumulative distribution function (cdf) of a standard normal

distribution $\mathcal{N}(0,1)$. Particularly, using the Barndorff-Nielsen approximation, the approximate cdf is of the form

$$\Phi\left\{c_1\sqrt{\mathrm{z_{LR}}} + \frac{1}{c_1\sqrt{\mathrm{z_{LR}}}}\log\frac{c\sqrt{\mathrm{z}_q}}{c_1\sqrt{\mathrm{z_{LR}}}}\right\}; \tag{2.8}$$

the formula of the Lugannani-Rice approximation is expressed as

$$\Phi(c_1\sqrt{\mathrm{z_{LR}}}) + \varphi(c_1\sqrt{\mathrm{z_{LR}}})\left(\frac{1}{c_1\sqrt{\mathrm{z_{LR}}}} - \frac{1}{c\sqrt{\mathrm{z}_q}}\right). \tag{2.9}$$

Here $\mathrm{z_{LR}}$ stands for the likelihood ratio test statistic value, z_q stands for the Wald or score test statistic value, Φ and φ denote the cdf and the probability density function of $\mathcal{N}(0,1)$, respectively, and $c_1 = \mathrm{sgn}(\hat{\beta} - \beta_0)$. Further, the equality $c = c_1$ holds for the Wald test statistic used in the approximation. Otherwise, using the score test statistic leads to the equality $c = \mathrm{sgn}(l'(\beta_0))$.

3. Comparison of tests

Test statistics are usually compared in terms of the size and power of tests. Let the null hypothesis $\mathrm{H}_0\colon\ \beta = \beta_0$ be tested on the significance level α. The size of a test corresponds to the probability of making a type I error, i.e., in the case of a two-sided test it is given as $\mathrm{P}(\mathrm{Z} \geqslant \chi^2_1(1-\alpha) \mid \mathrm{H}_0\colon\ \beta = \beta_0)$, or alternatively $\mathrm{P}(|\mathrm{Z}^*| \geqslant u(1-\frac{1}{2}\alpha) \mid \mathrm{H}_0\colon\ \beta = \beta_0)$, where Z or Z^* denotes the corresponding test statistics (2.2)–(2.7), $\chi^2_1(1-\alpha)$ is a $(1-\alpha)$-quantile of a chi-squared distribution with one degree of freedom and $u(1-\frac{1}{2}\alpha)$ represents a $(1-\frac{1}{2}\alpha)$-quantile of a standard normal distribution. The power of a two-sided test determined in a given point β^* is defined [16] as $\mathrm{P}(\mathrm{Z} \geqslant \chi^2_1(1-\alpha) \mid \mathrm{H_A}\colon\ \beta = \beta^*)$, or alternatively as $\mathrm{P}(|\mathrm{Z}^*| \geqslant u(1-\frac{1}{2}\alpha) \mid \mathrm{H_A}\colon\ \beta = \beta^*)$.

An asymptotic test property can cause disrespect of the nominal value α of the significance level for a small sample size. The tests can be either liberal, when the empirical size of the test (the rejection rate of true H_0) is higher than α, or conservative, when the empirical size of the test is smaller than α. For better comparability of the power of tests, it is therefore useful to consider the adjustment of the power. The *adjusted power* [30] of a two-sided test in a given point β^* is defined as $\mathrm{P}(\mathrm{Z} \geqslant (\chi^2_1(1-\alpha))^{\mathrm{adj}} \mid \mathrm{H_A}\colon\ \beta = \beta^*)$, differing from the power by the quantile $(\chi^2_1(1-\alpha))^{\mathrm{adj}}$, computed as a $(1-\alpha)$-quantile of the individual test statistic values calculated under the null hypothesis H_0. The main advantage of the adjusted power approach is the same empirical test size for all tests and approximations, which is equal to the nominal value α.

The theoretical power of three basic tests is based on the distribution of the test statistic under the alternative hypothesis $\mathrm{H_A}$. In our case, the likelihood ratio, the Wald test and the score test statistics [16] have asymptotically a noncentral chi-squared distribution with one degree of freedom under $\mathrm{H_A}$. The noncentrality parameter ψ can be determined using a Pitman-type alternative hypothesis [25], [29] $\mathrm{H_A}\colon\ \beta = \beta_0 + n^{-1/2}\Psi$, where $\Psi \in \mathbb{R}$. In general, the unknown parameter ψ can be expressed using the Fisher information $\mathcal{I}_1(\beta_0)$ of one observation as $\psi = \Psi^2\mathcal{I}_1(\beta_0)$. Applying the equality $n\mathcal{I}_1(\beta_0) = \mathcal{I}(\beta_0)$, where $\mathcal{I}(\beta_0)$ corresponds to the Fisher information of a random sample of size n, together with an expression of Ψ from the alternative hypothesis $\mathrm{H_A}$, we get the noncentrality parameter in the form of $\psi = (\beta - \beta_0)^2 \cdot \mathcal{I}(\beta_0)$.

The empirical size and power of tests can be determined as the proportion of rejection of the null hypothesis $\mathrm{H_0}$; true $\mathrm{H_0}$ in the case of the test size, false in the case of the power of the test. The corresponding adjustment of the empirical power is calculated similarly, except decision based on different critical values. Significant differences between test statistics in terms of the size or power of the test can be identified employing confidence intervals. These can be based on the confidence interval for a binomial proportion [2], [8]. The absolute frequency of rejection of the null hypothesis $\mathrm{H_0}$, denoted as a, follows a binomial distribution $\mathrm{Bi}(M, p)$, where M is the number of trials and p is the success probability for each trial. Obviously, the relative frequency of rejection equals $\hat{p} = a/M$. There are more possibilities [8] how to construct required confidence intervals, e.g. the exact confidence interval, or the normal approximation interval [2]. However, this approach brings problems related to the coverage probability for small sample sizes or the value of the proportion near to zero or one [8]. For this reason, the Wilson interval, with bounds equal to (see [2])

$$\left[\hat{p} + \frac{[u(1-\frac{1}{2}\alpha)]^2}{2M} \pm u\Big(1-\frac{1}{2}\alpha\Big)\sqrt{\frac{\hat{p}(1-\hat{p})}{M} + \frac{[u(1-\frac{1}{2}\alpha)]^2}{4M^2}}\right] \Big/ \Big(1 + \frac{[u(1-\frac{1}{2}\alpha)]^2}{M}\Big),$$

will be used.

4. Simulation study

Simulation and all computations were made by the statistical software R, version 3.6.1. The total number of 10,000 generated datasets M was considered and the sample size n of the individual datasets was set to 20, 50, 70, or 100. The probability of being censored was set gradually to 0%, 20%, or 50% for an observation.

The Weibull distribution $\mathrm{Wei}(\lambda, \nu)$ was used with survival times given as [5]

$$\text{survival time} = \Big(\frac{-\ln(U)}{\lambda \cdot \exp(x\beta)}\Big)^{1/\nu},$$

where $\lambda > 0$ represents a scale parameter, $\nu > 0$ stands for a shape parameter, U is a random variable from a uniform distribution $R(0, 1)$ and x is some covariate. Within the performed simulation study, the parameters $\lambda = 0.7$, or 1.7 and $\nu = 0.5$, 1, or 2 were considered. The parameter λ is directly related to the length of the survival time, i.e., $\lambda = 1.7$ represents longer survival time in comparison with $\lambda = 0.7$. Further, the parameter ν characterizes the behaviour of the hazard function. Namely, $\nu = 0.5$ corresponds to a decrease of the event hazard, $\nu = 1$ corresponds to a constant hazard function, and in the case of $\nu = 2$, the event hazard increases. The censoring times were generated from a uniform distribution $R(0, \theta)$, where θ was obtained from the equation [28]

$$\mathcal{P}(\delta = 0|\theta) = \int_D \mathcal{P}(\delta = 0 \mid x, \theta) f_x(x)\,\mathrm{d}x,$$

where $f_x(x)$ is the density function of the covariate x and D its domain. The probability $\mathcal{P}(\delta = 0 \mid \theta)$ of being censored was set to 0.2, or 0.5. The event times were considered without ties. One covariate x following either a standard normal distribution $\mathcal{N}(0, 1)$ or a binomial distribution $\mathrm{Bi}(n, 0.5)$ was examined. The null hypothesis $\mathrm{H}_0 \colon\ \beta = 0$ against the alternative $\mathrm{H}_\mathrm{A} \colon\ \beta \neq 0$ was tested on the 5% significance level.

The main aim of the simulation study is to investigate the behaviour of the size and power of three basic tests, i.e. the likelihood ratio (LiR), the Wald test (W) and the score test (S), as well as the approximations, i.e. the Barndorff-Nielsen (BNW and BNS, when the Wald and the score test is used, respectively) and the Lugannani-Rice (LRW, LRS) approximations. Concerning the size of a test, particular attention is paid to (i) the comparison of the size of tests with the nominal value α, (ii) the magnitude of the test size deviation from the nominal value α, and (iii) the behaviour under different settings (number of observations, probability of censoring, covariate distribution, and type of the event hazard and the length of survival time). In the case of the power of a test, the study is mainly focused on (i) the behaviour of adjusted power curves under different settings, (ii) the behaviour of adjusted power curves in the neighbourhood of the point $\beta = 0$, (iii) how fast the tests achieved a sufficiently large adjusted power of 0.8, and (iv) the comparison of adjusted power curves with a theoretical power curve for basic tests. Due to the symmetry of a power curve with respect to the $\beta = 0$ axis, all power curves are plotted only for the nonnegative values of β.

Based on the performed simulations, let us summarize the main results. First, the size of tests will be discussed. The statistical significance of the deviation between the test size and the nominal value 5% was assessed using the 95% confidence intervals. The relative frequency of cases when tests have been identified as liberal (the confidence interval is above the nominal value) or conservative (the confidence interval is below the nominal value) in all 72 configurations is presented in Table 1. For both types of the covariate distribution, situations when the tests were liberal prevailed. The Wald test was conservative in a few cases for both distributions. Other tests were conservative only once for the discrete covariate. In general, the Wald test was identified as the test with the most accurate test size (Table 1, Figure 1). In contrast to the other tests, that accuracy was emphasized even for a small sample size ($n = 20$) or higher probability of censoring, mainly for the continuous covariate. However, we cannot state in which situations the Wald test is liberal in general.

x	test	LiR	W	S	BNW	BNS	LRW	LRS
$\mathcal{N}(0,1)$	liberal	0.806	0.556	0.806	0.806	0.819	0.806	0.819
	accurate	0.194	0.417	0.194	0.194	0.181	0.194	0.181
	conservative	0	0.028	0	0	0	0	0
Bi$(n, 0.5)$	liberal	0.667	0.486	0.819	0.639	0.722	0.639	0.722
	accurate	0.319	0.444	0.167	0.347	0.264	0.347	0.264
	conservative	0.014	0.069	0.014	0.014	0.014	0.014	0.014

Table 1. The relative frequency of liberality and conservativeness of LiR, W, S, BNW, BNS, LRW, and LRS.

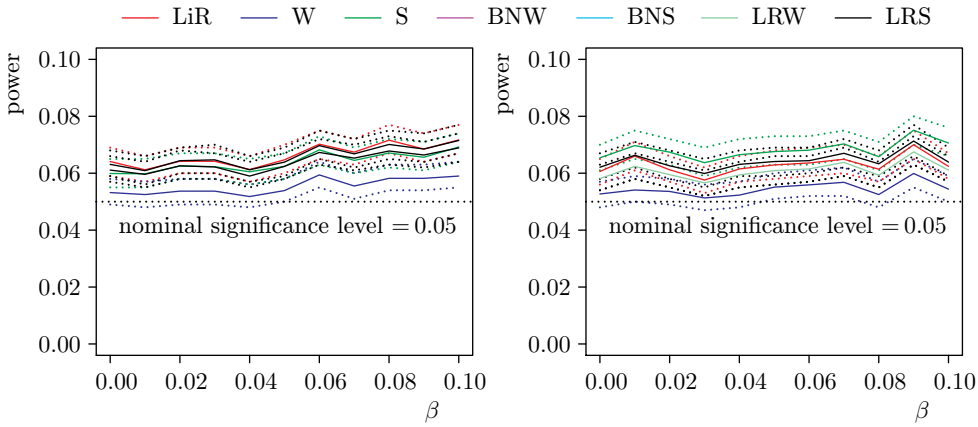

Figure 1. The empirical power curves (solid lines) together with the pointwise 95% confidence bands (dotted lines) in the case of a continuous (left) and a discrete (right) covariate. Settings: $n = 20$, $\lambda = 0.7$, $\nu = 1$, 0.2 probability of censoring.

Considering various settings of the simulation in more detail, there were 74% and 64% cases of equality between the decision about the test size for the continuous and for the discrete covariate, respectively. The increasing number of observations resulted in a shift of the tests from liberal to accurate, except the predominant liberality of the Wald test for a continuous covariate. The shift was more evident for the discrete covariate. A higher probability of censoring implied a higher accuracy of the test size of the Wald test for both types of the covariate and also of the BN and the LR approximations of the Wald test for a discrete covariate only. Unfortunately, we cannot state in general that the size of approximations is more accurate than that of the basic tests. We only found that the approximations of the score test were identified as more accurate than the score test from 50 observations for a discrete covariate.

Let us look at the influence of the length of the survival time (λ) and the type of the hazard function (ν) on the size of tests. For an increasing survival time (λ increasing), the size of all tests does not change significantly with a few exceptions for the Wald test (three times for a discrete covariate for $n = 20$ or 50, twice for a continuous covariate for $n = 20$, all under various settings of other parameters). Further, with the change of the hazard function from decreasing through constant to increasing (the value of ν increasing), the liberality of tests rose for both distributions of the covariate. The effect was more distinctive for a continuous covariate (Table 2). There was only one case when the Wald test was identified as conservative for an increasing hazard function ($n = 20$, $\lambda = 0.7$, 0.5 probability of censoring).

ν	test	LiR	W	S	BNW	BNS	LRW	LRS
0.5	liberal	0.667	0.458	0.667	0.667	0.667	0.667	0.667
	accurate	0.333	0.542	0.333	0.333	0.333	0.333	0.333
	conservative	0	0	0	0	0	0	0
2	liberal	0.917	0.583	0.917	0.917	0.917	0.917	0.917
	accurate	0.083	0.375	0.083	0.083	0.083	0.083	0.083
	conservative	0	0.042	0	0	0	0	0

Table 2. The relative frequency of liberality and conservativeness of LiR, W, S, BNW, BNS, LRW, and LRS in the case of a continuous covariate.

Finally, the magnitude of the deviation from the nominal size of the test will be investigated. Let us define a test as *weakly liberal/conservative* if the empirical size of the test is below/above the value accounting for 1.5/0.5 times the nominal significance level, i.e., the 95% confidence interval for the size of the test does not exceed $[0.07, 0.08]$/drop below $[0.02, 0.03]$. For both types of covariate, all liberal and conservative tests were marked as weakly liberal and weakly conservative, respectively,

with a few exceptions for a discrete covariate. Particularly, the Wald test was not weakly conservative in the case of the sample size $n = 20$, $\lambda = 0.7$, $\nu = 0.5$ and the probability of censoring 0.5. The score test was three times identified as not weakly liberal, all with the sample size $n = 20$.

Second, the power of the tests will be explored. Comparison of the empirical power curves was based on the idea of adjusted power together with the corresponding pointwise 95% confidence bands. The individual empirical (adjusted) power curves were situated lower, i.e., tests were less powerful, for a higher probability of censoring because of fewer observations which were taken into account. They shrunk and were approaching the theoretical power curve with an increasing number of observations. Moreover, an increasing number of observations led to the mutual similarity between the individual empirical power curves.

Let us examine the effect of the length of survival time (λ) and the type of event hazard (ν) on the adjusted power of tests. In the case of a shorter survival time ($\lambda = 0.7$), the effect of the change of the event hazard (ν) on the adjusted power of tests was identified for an arbitrary sample size and 0.5 probability of censoring. Particularly, the change of the hazard function from decreasing through constant to increasing (ν increasing) led to less and more powerful tests for a continuous and discrete covariate, respectively (Figure 2). For other settings, when the probability of censoring was zero or 0.2, or for a longer survival time ($\lambda = 1.7$), these effects did not occur.

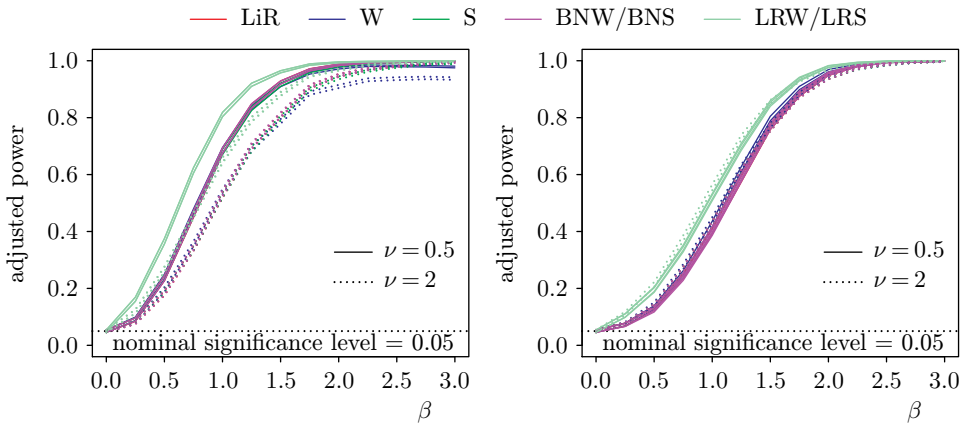

Figure 2. The pointwise 95% confidence bands for the adjusted power curves in the case of a continuous (left) and a discrete (right) covariate. Settings: $n = 20$, $\lambda = 0.7$, 0.5 probability of censoring.

Moreover, the effect of the length of survival time (λ) for a fixed event hazard (ν) and an arbitrary sample size was examined. For zero probability of censoring, the adjusted empirical power did not change in relation to the considered survival time (λ).

For a higher probability of censoring (0.2 or 0.5), the adjusted power was increasing for a longer survival time (λ increasing). Focusing on various values of the event hazard (ν), for an increasing ν, there was a decrease in the difference in adjusted power curves for the considered length of survival time in general. There is one exception for a continuous covariate with 0.5 probability of censoring; the difference was increasing together with an increase of the event hazard (ν increasing), see Figure 3.

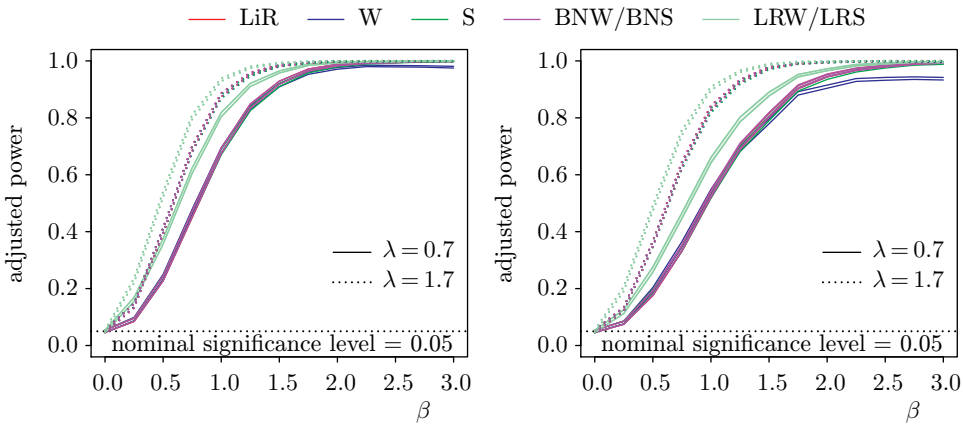

Figure 3. The pointwise 95% confidence bands for the adjusted power curves in the case of a continuous covariate. Settings: $n = 20$, $\nu = 0.5$ (left) and $\nu = 2$ (right), 0.5 probability of censoring.

Focusing on the neighbourhood of $\beta = 0$, i.e. from 0 to 0.6, in all considered variants, the LR approximations were significantly more powerful than all the basic tests and the BN approximations, except overlapping in β around the value 0.1 (mainly for a discrete covariate) or in higher values of β, i.e. 0.5 or 0.6 ($n = 100$, a continuous covariate), see Figures 2, 3, 4. There is no significant difference in the adjusted power between the LRS and the LRW. Further, there is no difference in the adjusted power between the basic tests and the BN approximations. Here, only two exceptions were identified, when the Wald test was significantly more powerful than the likelihood ratio test in β around 0.6 (a discrete covariate, 0.5 probability of censoring).

Next, let us look at how fast the tests achieved a sufficiently large adjusted power of 0.8 (Figures 2, 3). The value of β, where the required power of 0.8 was achieved, was determined by cubic interpolation of the upper bound of the pointwise confidence band for the adjusted power. Then, the mean of β and the standard error over the various situations with the same sample size were calculated (Table 3). In general, the three basic tests and the corresponding BN approximations reached the required power for almost the same values of β for individual simulation setup. These

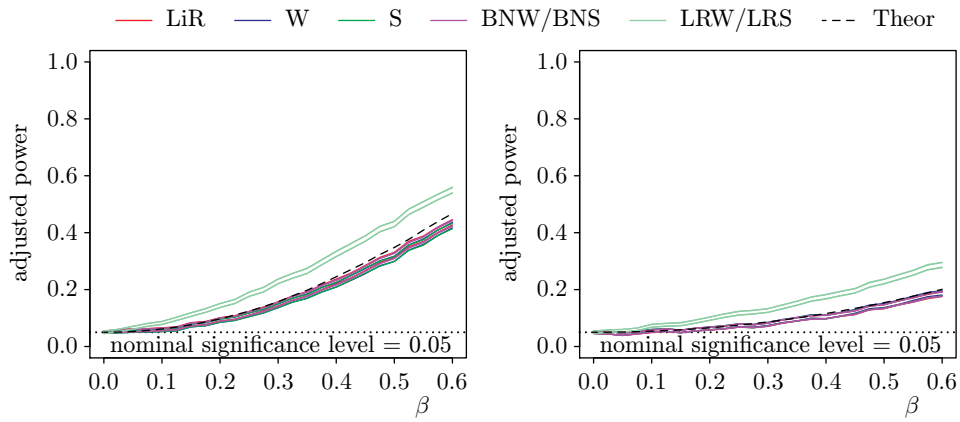

Figure 4. The pointwise 95% confidence bands for the adjusted power curves in the case of a continuous (left) and a discrete (right) covariate. Settings: $n = 20$, $\lambda = 0.7$, $\nu = 1$, 0.2 probability of censoring.

values of β were significantly higher than those of the corresponding LR approximations for the discrete covariate. That emphasized higher adjusted power of the LR approximations. For the continuous covariate, a significant difference of the values β was found for sample sizes 20 and 50, only. In addition to that, the adjusted power of the tests and the approximations was higher for the continuous covariate than for the discrete one. A similar pattern occurred also for various types of the hazard function considered separately.

n	x	LiR	W	S	BNW	BNS	LRW	LRS
20	$\mathcal{N}(0,1)$	0.935	0.941	0.942	0.935	0.936	0.804	0.805
		(0.043)	(0.045)	(0.044)	(0.043)	(0.043)	(0.037)	(0.037)
	Bi$(n, 0.5)$	1.531	1.528	1.537	1.529	1.532	1.318	1.319
		(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.008)
100	$\mathcal{N}(0,1)$	0.379	0.379	0.379	0.379	0.379	0.328	0.327
		(0.014)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
	Bi$(n, 0.5)$	0.623	0.622	0.622	0.622	0.622	0.549	0.549
		(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.011)	(0.011)

Table 3. The mean and the standard error of the values of β, at which the adjusted power of 0.8 was achieved for LiR, W, S, BNW, BNS, LRW, and LRS.

In the neighbourhood of β with average power 0.8, the adjusted power of the LR approximations was significantly higher than that of the other tests and the BN approximations for both types of covariate (Figures 2, 3). For a continuous covariate, the likelihood ratio, the Wald test, the score test and the BN approximations over-

lapped. However, there were more obvious differences between these tests in a discrete case. Namely, there were four cases (small sample size, or higher probability of censoring) when the Wald test had higher adjusted power than the score test and the BN approximations.

x	test	LiR	W	S	BNW	BNS	LRW	LRS
$\mathcal{N}(0,1)$	more powerful	0.181	0.153	0.153	0.181	0.181	0.986	0.986
	similar	0.514	0.528	0.528	0.514	0.514	0.014	0.014
	less powerful	0.306	0.319	0.319	0.306	0.306	0	0
$\mathrm{Bi}(n,0.5)$	more powerful	0.972	0.986	0.986	0.972	0.972	1	1
	similar	0.028	0.014	0.014	0.028	0.028	0	0
	less powerful	0	0	0	0	0	0	0

Table 4. The relative frequency of the similarity of the adjusted power for LiR, W, S, BNW, BNS, LRW, and LRS to the theoretical power of 0.8.

ν	test	LiR	W	S	BNW	BNS	LRW	LRS
0.5	more powerful	0.25	0.25	0.25	0.25	0.25	1	1
	similar	0.5	0.5	0.5	0.5	0.5	0	0
	less powerful	0.25	0.25	0.25	0.25	0.25	0	0
2	more powerful	0.083	0.042	0.042	0.083	0.083	0.958	0.958
	similar	0.542	0.542	0.542	0.542	0.542	0.042	0.042
	less powerful	0.375	0.417	0.417	0.375	0.375	0	0

Table 5. The relative frequency of similarity of the adjusted power for LiR, W, S, BNW, BNS, LRW, and LRS to the theoretical power of 0.8 in the case of the continuous covariate.

Finally, a comparison of the theoretical and adjusted power curves for all tests was performed. With the usage of cubic interpolation, the values of β corresponding to the theoretical power of 0.8 were determined. In the discrete case, the adjusted power of all tests was mostly significantly higher than the theoretical one and only about 0–2% from the total number of the variants were similar (Table 4). A similar pattern has also occurred for a variety of the hazard functions assessed separately. On the other hand, there was an obvious difference in the behaviour of the LR approximations and the other tests for a continuous covariate (Tables 4, 5). Namely, for the other tests, the adjusted power of about 15–20% of cases was significantly higher than the theoretical one, while for the LR approximations it was almost all the cases. About 50% of variants were determined as similar to the theoretical power, contrary to the LR approximations with only about 1%. In contrast to the discrete covariate, there was about 30% of cases with significantly lower adjusted power than

the theoretical power, while for the LR approximations these was none. The change of the hazard function from increasing through constant to decreasing (the value of ν is decreasing) resulted mostly in a higher proportion of tests considered as more powerful than the theoretical power curve. That highlighted the usage of LR approximations for both discrete and continuous covariate.

5. Conclusions and discussion

The likelihood ratio, the Wald test and the score test, together with the Barndorff-Nielsen and the Lugannani-Rice approximations were examined in terms of the test size and power in the Cox proportional hazards model with a simple covariate under different settings (sample size $n = 20$, 50, 70, 100; probability of censoring set to 0, 0.2, or 0.5; a covariate with a binominal and a standard normal distribution; a Weibull distribution $\mathrm{Wei}(\lambda, \nu)$ of a baseline hazard function with $\lambda = 0.7$, 1.7 and $\nu = 0.5$, 1, 2). A two-sided test with the null hypothesis $\mathrm{H}_0\colon\ \beta = 0$ was tested on the significance level $\alpha = 0.05$.

Taking into account the test size, the empirical test sizes were approaching the nominal value with an increasing number of observations. In general, a weak liberality (the empirical test size not greater than 0.075) prevailed for all tests, with a few exceptions for a discrete covariate. The Wald test came out as the test with the most accurate test size for both distributions of the covariate, even for situations with only 20 observations, or with a higher probability of censoring. An increasing sample size led to more accurate test size for all tests, except the Wald test with the most accurate test size in general. An increase in the probability of censoring implied a higher accuracy of the Barndorff-Nielsen and the Lugannani-Rice approximations of the Wald test for both distributions of the covariate. Finally, a change of event hazard in terms of increasing values of ν led to more liberal tests, mainly for a continuous covariate.

Based on the adjusted power, an increasing number of observations led to the mutual similarity of the individual adjusted power curves and also the theoretical one. An increasing proportion of censored data implied a decrease in the adjusted power. The Lugannani-Rice approximations were more powerful than the other tests for both considered distributions, with one exception ($n = 100$, a continuous covariate). For a shorter length of survival time ($\lambda = 0.7$) and 0.5 probability of censoring, a change of event hazard in terms of increasing values of ν led to a slower and a faster rise in the adjusted power for a continuous and a discrete covariate, respectively.

In practice, only the most common type of null hypothesis is verified. If $\beta = 0$ (or, equivalently, the hazard ratio $\exp(\beta) = 1$), the covariate does not affect the event hazard. A value β greater than zero (a hazard ratio greater than one) indicates

that as the value of the covariate increases, the event hazard increases and thus the length of survival time decreases. Similarly, a value β less than zero (a hazard ratio less than one) indicates that as the value of the covariate increases, the event hazard reduces and thus the length of survival time increases. From a theoretical point of view, different values of β_0 can be assumed as well and the behaviour of the size and power of tests can be analogously investigated. The behaviour of the power curves will be similar, up to a shift of the power curves in a horizontal direction, and a change of their slope.

In addition to the analyzed distributions of a covariate, other alternatives, e.g. a trimmed or skewed normal distribution or some bimodal distribution could be considered. However, for the two distributions considered in this paper and for the three and two settings of the event hazard (ν) and the length of survival time (λ), respectively, the study showed many patterns in the behaviour of the tests. Moreover, the simulation could be enriched by a larger number of covariates. The need for the extension of this concept was also outlined in [4].

A c k n o w l e d g e m e n t. The authors thank the referee for useful comments and suggestions, which helped significantly improve the paper. We also want to thank Silvie Bělašková, PhD for the discussion of higher-order approximations and Jan Elgner for valuable advice with the statistical software R.

References

[1] *A. O. Adujemo, A. O. Ahmadu*: A study of the slope of Cox proportional hazard and Weibull models: Simulated and real life data approach. Science World Journal *11* (2016), 31–35.

[2] *A. Agresti, B. A. Coull*: Approximate is better than "exact" for interval estimation of binomial proportions. Am. Stat. *52* (1998), 119–126. MR doi

[3] *O. Barndorff-Nielsen, D. R. Cox*: Edgeworth and saddle-point approximations with statistical applications. J. R. Stat. Soc., Ser. B *41* (1979), 279–312. zbl MR doi

[4] *S. Bělašková, E. Fišerová*: Improvement of the accuracy in testing the effect in the Cox proportional hazards model using higher order approximations. Filomat *31* (2017), 5591–5601. MR doi

[5] *R. Bender, T. Augustin, M. Blettner*: Generating survival times to simulate Cox proportional hazards models. Stat. Med. *24* (2005), 1713–1723. MR doi

[6] *A. R. Brazzale, M. Valentina*: Likelihood asymptotics in nonregular settings: A review with emphasis on the likelihood ratio. Working Paper Series *4* (2018), 45 pages; Available at `http://paduaresearch.cab.unipd.it/11306/`.

[7] *N. Breslow*: Discussion of Professor Cox's paper. J. R. Stat. Soc., Ser. B *34* (1972), 216–217. zbl MR

[8] *L. D. Brown, T. T. Cai, A. DasGupta*: Interval estimation for a binomial proportion. Stat. Sci. *16* (2001), 101–133. zbl MR doi

[9] *A. Buse*: The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. Am. Stat. *36* (1982), 153–157. doi

[10] *T. K. Chandra, S. N. Joshi*: Comparison of likelihood ratio, Rao's and Wald's tests and a conjecture of C. R. Rao. Sankhyā, Ser. A *45* (1983), 226–246. zbl MR
[11] *D. R. Cox, D. Oakes*: Analysis of Survival Data. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1984. MR doi
[12] *A. M. Crumer*: Comparison Between Weibull and Cox Proportional Hazards Models. Kansas State University, Manhattan, 2011; Available at `https://core.ac.uk/download/pdf/5172563.pdf`.
[13] *B. Efron*: The efficiency of Cox's likelihood function for censored data. J. Am. Stat. Assoc. *72* (1977), 557–565. zbl MR doi
[14] *E. Fišerová, M. Chvosteková, S. Bělašková, M. Bumbálek, Z. Joska*: Survival analysis of factors influencing cyclic fatigue of nickel-titanium endodontic instruments. Adv. Mater. Sci. Engineer. *2015* (2015), Article ID 189703, 6 pages. doi
[15] *D. A. S. Fraser, J. Wu, A. C. M. Wong*: An approximation for noncentral chi-squared distribution. Commun. Stat., Simulation Comput. *27* (1998), 275–287. zbl MR doi
[16] *D. W. Gudicha, V. D. Schmittmann, J. K. Vermunt*: Statistical power of likelihood ratio and Wald test in latent class models with covariates. Behavior Research Methods *49* (2017), 1824–1837. doi
[17] *D. W. Hosmer,* Jr., *S. Lemeshow*: Applied Survival Analysis: Regression Modeling of Time to Event Data. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, 1999. zbl MR doi
[18] *A. Ihwah*: The use of Cox regression model to analyze the factors that influence consumer purchase decision on a product. Agriculture and Agricultural Science Procedia *3* (2015), 78–83. doi
[19] *J. F. Lawless*: Statistical Models and Methods for Lifetime Data. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, 2003. zbl MR doi
[20] *E. T. Lee, O. T. Go*: Survival analysis in public health research. Annu. Rev. Public Health *18* (1997), 105–134. doi
[21] *R. Lugannani, S. Rice*: Saddle point approximation for the distribution of the sum of independent random variables. Adv. Appl. Probab. *12* (1980), 475–490. zbl MR doi
[22] *D. A. Pierce, R. Bellio*: Beyond first-order asymptotics for Cox regression. Bernoulli *21* (2015), 401–419. zbl MR doi
[23] *J. Qian, B. Li, P. Chen*: Generating survival data in the simulation studies of Cox model. Third International Conference on Information and Computing. IEEE, Los Alamitos, 2010, pp. 93–96. doi
[24] *M. Schemper*: Cox analysis of survival data with non-proportional hazard functions. J. R. Stat. Soc., Ser. D *41* (1992), 455–465. doi
[25] *P. K. Sen, J. M. Singer*: Large Sample Methods in Statistics: An Introduction With Applications. Chapman & Hall, New York, 1993. zbl MR doi
[26] *I. M. Skovgaard*: An explicit large-deviation approximation to one-parameter tests. Bernoulli *2* (1996), 145–165. zbl MR doi
[27] *I. M. Skovgaard*: Likelihood asymptotics. Scand. J. Stat. *28* (2001), 3–32. zbl MR doi
[28] *F. Wan*: Simulating survival data with predefined censoring rates for proportional hazards models. Stat. Med. *36* (2017), 838–854. MR doi
[29] *Y. Yi, X. Wang*: Comparison of Wald, score, and likelihood ratio tests for response adaptive designs. J. Stat. Theory Appl. *10* (2011), 553–569. MR
[30] *J. Zhang, D. D. Boos*: Adjusted power estimates in Monte Carlo experiments. Commun. Stat., Simulation Comput. *23* (1994), 165–173. zbl doi
[31] *J. Zhang, J. E. Kolassa*: A comparison of the accuracy of saddlepoint conditional cumulative distribution function approximations. Complex Datasets and Inverse Problems:

Tomography, Networks and Beyond. IMS Lecture Notes Monograph Series 54, Institute of Mathematical Statistics, Beachwood, 2007, pp. 250–259. MR doi

Authors' address: *Aneta Andrášiková* (corresponding author), *Eva Fišerová*, Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc, 17. listopadu 12, 771 46 Olomouc, Czech Republic, e-mail: aneta.andrasikova@upol.cz, eva.fiserova@upol.cz.