Filomat 31:18 (2017), 5591–5601
https://doi.org/10.2298/FIL1718591B

Published by Faculty of Sciences and Mathematics,
University of Niš, Serbia
Available at: http://www.pmf.ni.ac.rs/filomat

Improvement of the Accuracy in Testing the Effect in the Cox Proportional Hazards Model Using Higher Order Approximations

Silvie Bělašková[a], Eva Fišerová[b]

[a] *International Clinical Research Center, St. Anne's University Hospital Brno, Pekařská 53, 656 91 Brno, Czech Republic*
[b] *Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University in Olomouc, 17.listopadu 12, 771 46 Olomouc, Czech Republic*

Abstract. Small-sample properties of the likelihood ratio test, the Wald test and the score test about significance of the effect in the Cox proportional hazards model for the right-censored and left-truncated data are investigated. These are large-sample tests, and, therefore, these are only approximate tests and they do not necessary maintain chosen significance level. Consequently, the p-value can be inaccurate as well. Higher order approximations of the likelihood function based on the Barndorff-Nielsen formula and the Lugannani-Rice formula are used in order to improve the accuracy of statistical inferences. The accuracy of these tests together with proposed approximations are compared by means of simulations under conditions of decreasing the sample size, and increasing proportion of right-censored and left-truncated data in the Cox model with the exponential and the Weibull distribution of the baseline hazard function. The results show that higher order approximations based on the Lugannani-Rice and the Barndorff-Nielsen formulas in combination with the Wald statistic improve the accuracy.

1. Introduction

Time to event data analysis, called as survival analysis, is in medicine one of the most used approach at all. It is a collection of statistical procedures for analysing the duration until the occurrence of an event of interest. An event is usually taken as death, relapse of some disease, stage of disease, etc. The methodology is also applicable in other fields like in economy [4, 30], material technology [19], or information technology [25]. Observed times of event are usually analysed by means of the survival analysis, however other basic approaches like a linear model approach or a comparative analysis can be used.

A typical phenomenon for survival analysis is a censoring and a truncation. A censoring is used when missing information about the time of the event occurs. For example, an event happened in some time interval, did not occur during the study, or an individual left the study before its end. The last two examples represent a right-censoring and in our study the latest one is considered. Left-truncation, also called as delayed entry, means that the start of follow up for some individuals is different from the specified time origin, so these individuals are observed after they survive some entry point.

2010 *Mathematics Subject Classification*. Primary 62N03; Secondary 62N01
Keywords. Higher order asymptotics; likelihood function; likelihood ratio test; Wald test; score test; Cox proportional hazards model; Lugannani-Rice formula; Barndorff-Nielsen formula; censoring; truncation.
Received: 11 November 2016; Accepted: 21 December 2016
Communicated by Mića Stanković
Research supported by the project no. LQ1605 from the National Program of Sustainability II (MEYS CR)
Email addresses: silvie.belaskova@fnusa.cz (Silvie Bělašková), eva.fiserova@upol.cz (Eva Fišerová)

In time to event data analysis, we have to select a time scale and specify the origin and the end of the study. A time scale is usually time-on-study and the age is considered as a covariate. Recently, the age as a time scale started to be considered as well [36]. When age is considered as a time scale, the beginning of these time to event studies is the date of the birth of studied individuals. These data are considered like left-truncated [14]. There is amount of scientific articles focused on choosing a time scale and defining starting point, e.g. [21, 24, 29]. Meanwhile Korn et al. [29] prefers age of the patience as the time scale, Ingram and Makuc [24] suggested two easy conditions which provides that time to study do not give biased estimate in case when age is used as a time scale. Authors Gail et al. [21] focused on six different Cox proportional hazards models with respect to the time scale used. Left-truncation is in the medical field used not just for age as a time scale but also when time to event after some specific point is analysed, e.g. twenty-four hours after a surgery, or seven days after leaving a hospital.

Common statistical task in survival analysis is to model the effect of considered covariates on time to event. There are several parametric, semi-parametric and non-parametric models in survival analysis which can be used to describe time to event, see e.g. [22]. The Cox proportional hazards model [15] is one of the most used model in survival analysis and is classified as a semi-parametric model.

The aim of the paper is to evaluate the accuracy of statistical inference about a scalar parameter in the Cox proportional hazards model. The significance of the effect of each covariate in the Cox model is usually verified by means of the likelihood ratio test, the Wald test and the score test [11]. These are large-sample tests, and therefore, these are only approximate tests and they do not necessary maintain the significance level α. Accordingly, higher order approximations [8] of the likelihood root based on the Barndorff-Nielsen formula [3] and the Lugannani-Rice formula [33] are used in order to improve the accuracy of statistical inferences. The accuracy of these large-sample tests together with the proposed approximations are compared by means of simulations under conditions of decreasing the sample size, and increasing proportion of right-censored and left-truncated data in the Cox model with the exponential and the Weibull distribution of the baseline hazard function.

The paper is organized as follow. In the next section, some fundamentals of the Cox proportional hazards model, necessary for this contribution, are recalled. Section 3 is devoted to the large-sample tests for testing significance of the effect of one covariate in the Cox model. The accuracy of tests is discussed in the context of liberal and conservative tests. In Section 4, higher order approximations are introduces. A large simulation study focused on the accuracy of the tests is presented in Section 5 and the final Section 6 concludes.

2. Cox Proportional Hazards Model

The widely used method to investigate several variables at a time is the Cox proportional hazards model [15]. The hazard function at a time t is the probability that, during a very short time interval, an event will occur, conditional on not having the event up to a time t. The hazard function assesses the instantaneous risk of the event for an individual which has survived a time t. Formally, let $T \geq 0$ be a random variable denoting the event time. The hazard function is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P\{t \leq T < t + \Delta t | T \geq t\}}{\Delta t}. \tag{1}$$

The Cox proportional hazards model specifies the hazard function to covariates $\mathbf{x} = (x_1, \ldots, x_p)'$ for an i-th individual in the form

$$h(t, \mathbf{x}_i) = h_0(t) \exp(x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ip}\beta_p), \tag{2}$$

where $h_0(t)$ is the baseline hazard function when all covariates are zero. The key assumption in the Cox models is proportional hazards, i.e. the hazard for any individual is a fixed proportion of the hazard for any other individual, which can be limiting in some cases. However, the advantage of the Cox model is that no assumptions are made about the baseline hazard and it is not necessarily to be estimated.

With respect to the proportionality of the hazard function, Cox in [15] suggested the partial likelihood function for the estimation of regression coefficients. This approach is based on the part of the likelihood function representing the main effect of the covariates which is free of the baseline hazard. Assuming that event times and censoring times are independent, the partial likelihood is the product of every conditional probabilities of an event of the i-th individual at time t_i given events of individuals at risk at time t_i, that is

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}_i'\beta)}{\sum_{j \in R_i} \exp(\mathbf{x}_j'\beta)} \right)^{\delta_i}. \tag{3}$$

Here the symbol $R_i = \{j : t_j \geq t_i\}$ denotes the risk set comprising those individuals still available to have an event at time t_i and t_j means the event time of those subjects at risk. Using the risk set is a convenient mechanism for excluding from denominator those individuals who already experienced the event and from this point of view are not part of this risk set [1]. The symbol δ_i is an indicator of the survival status of the i-th individual, where $\delta_i = 1$ denotes an event at time t_i and $\delta_i = 0$ means right-censoring.

The corresponding log partial likelihood is

$$l(\beta) = \sum_{i=1}^{n} \left(\mathbf{x}_i'\beta - \log \sum_{j \in R_i} \exp(\mathbf{x}_j'\beta) \right)^{\delta_i}, \tag{4}$$

the partial score function is

$$l'(\beta) = \sum_{i=1}^{n} \left(\mathbf{x}_i - \frac{\sum_{j \in R_i} \exp(\mathbf{x}_j'\beta)\mathbf{x}_j}{\sum_{j \in R_i} \exp(\mathbf{x}_j'\beta)} \right)^{\delta_i}, \tag{5}$$

and the Hessian matrix of the partial log likelihood is

$$l''(\beta) = -\sum_{i=1}^{n} \left(\frac{\sum_{j \in R_i} \exp(\mathbf{x}_j'\beta)\mathbf{x}_j\mathbf{x}_j'}{\sum_{j \in R_i} \exp(\mathbf{x}_j'\beta)} - \frac{[\sum_{j \in R_i} \exp(\mathbf{x}_j'\beta)\mathbf{x}_j][\sum_{j \in R_i} \exp(\mathbf{x}_j'\beta)\mathbf{x}_j']}{[\sum_{j \in R_i} \exp(\mathbf{x}_j'\beta)]^2} \right)^{\delta_i}. \tag{6}$$

The inverse of the Hessian matrix evaluated at the maximum likelihood estimate $\widehat{\beta}$ is often used as an approximate variance-covariance matrix for the estimator $\widehat{\beta}$. The negative value of the Hessian matrix is the observed Fisher information matrix, $\mathbf{J}(\widehat{\boldsymbol{\beta}})$ [31].

Handling with left-truncation is performed by control of the risk set R_i. Let us denote a truncation time as K. Then the risk set of the partial likelihood function is given by $R_i = \{j : t_j \geq t_i \wedge t_j \geq K\}$.

When times in the continuous time model are grouped, ties in event times can be observed. In the Cox partial likelihood model ties are not considered, because the formula (3) is valid only for data which are not grouped. If the number of tied data is relatively small, Breslow [10] proposed an approximation of the likelihood function based on summing up covariates for all individuals experiencing the event at a given time point t_i and rising the result to a power equal to the number of events tied at t_i. In case that no tied data occur, the Breslow approximation gives the same results as the Cox model.

3. Testing Significance of the Effect

The significance of the effect of each covariate in the Cox proportional hazards model is usually verified by means of the likelihood ratio test, the Wald test and the score test [11]. In view of issues in the following sections, only one covariate x will be considered in the Cox model. However, all the results can be generalized for a p-dimensional vector of covariates. For a scalar regression parameter β, these tests are based on the functions

$$\text{likelihood root} \quad r(\beta) = \text{sign}(\widehat{\beta} - \beta)\left[2\left\{l(\hat{\beta}) - l(\beta)\right\}\right]^{1/2}; \tag{7}$$

$$\text{score statistic} \quad s(\beta) = j(\hat{\beta})^{-1/2}\partial l(\beta)/\partial\beta; \tag{8}$$

$$\text{Wald statistic} \quad t(\beta) = j(\widehat{\beta})^{-1/2}(\widehat{\beta} - \beta), \tag{9}$$

where $j(\widehat{\beta})$ is the observed Fisher information. Under the null hypothesis $H_0 : \beta = \beta_0$, the statistics (7)-(9) have an asymptotic standard normal distribution, $N(0,1)$. In survival analysis, these statistics are used in a square. Hence, under the null hypothesis, the likelihood ratio test statistic is

$$T_{LRT} = 2[l(\widehat{\beta}) - l(\beta_0)], \tag{10}$$

the Wald test is based on a statistic of the form

$$T_W = (\widehat{\beta} - \beta_0)^2 \Big/ j(\widehat{\beta}), \tag{11}$$

and the test statistic for the score test is

$$T_S = [\partial l(\beta_0)/\partial\beta]^2 \Big/ j(\widehat{\beta}). \tag{12}$$

The statistics (10)-(12) have an asymptotic χ^2-distribution with one degree of freedom if the null hypothesis is true. For a given significance level α, H_0 is rejected if the realization of the test statistic is greater than the $(1-\alpha)$-quantile of χ^2-distribution with one degree of freedom. These are large-sample tests. Therefore, these are only approximate tests and they do not necessary maintain the significance level α.

The decision to reject or accept a null hypothesis is mostly based on a p-value. Small values of the p-value $p(\mathbf{X})$ give evidence that an alternative hypothesis is true. The p-value is said to be valid if $P(p(\mathbf{X}) \leq \alpha | H_0) \leq \alpha$ for every $\alpha \in \langle 0,1\rangle$ [7]. If the test is approximate, the p-value can be inaccurate. It is possible that the observed p-value p is smaller than the true p-value is. A test that tends to underestimate the true p-value is called a liberal test. A liberal test is more likely to find statistical significance even where it does not truly exist. In contrast, the test is conservative if it tends to overestimate the p-value p. Hence, the probability of incorrectly rejecting the null hypothesis is never greater than the significance level α. A conservative test is less likely to find statistical significance even where it does truly exist.

The valid p-value $p(\mathbf{X})$ follows the uniform distribution $\mathcal{U}(0,1)$ under the null hypothesis, however this not happened generally in the case of approximate tests. To assess the accuracy of the p-value, the empirical cumulative distribution function of $p(\mathbf{X})$ can be compared with the theoretical cumulative distribution function of $\mathcal{U}(0,1)$ [37]. As it will be shown later, the accuracy of the p-value depends not only on the sample size, but also on the proportion of right-censored observations in a data set and on the length of left-truncation as well.

4. Higher Order Approximations

Asymptotically, the likelihood ratio test, the Wald test and the score test have the same distribution, however, numerically they give different results in applications. Numerous books and papers deal with properties of these test. Peers [34] showed that under general conditions there is no huge difference between these tests from the power of these functions side. Chandra and Joshi [12] proved that the score test is more powerful than these others two for the large-sample size. Yi [40] compared these tests under a specific design of experiment and based on the simulation study recommends to use the Wald test.

These three tests are approximations of the first order with the relative error of order $O(n^{-1/2})$. In the following, modification of the likelihood root based on the theory of higher order asymptotics [8] is used in order to improve the accuracy of the p-value when testing the significance of an effect in the Cox proportional hazard model with one covariate. The Barndorff-Nielsen approximation [3] modifies the likelihood root by adding the term with the combination of the likelihood root with the score statistic or the Wald statistic as follows

$$r^*(\beta) = r(\beta) + \frac{1}{r(\beta)} \log\left(\frac{q(\beta)}{r(\beta)}\right), \tag{13}$$

where $q(\beta) = s(\beta)$ (the combination of the likelihood root and the score statistic), or $q(\beta) = t(\beta)$ (the combination of the likelihood root and the Wald statistic). The statistic r^* has an asymptotic standard normal

distribution, or, equivalently, $(r^*)^2$ has an asymptotic χ^2-distribution with one degree of freedom. Other type of modification is the Lugannani-Rice approximation [33]. Its distribution function has similar asymptotic properties like the distribution function of r^*. The Lugannani-Rice formula is defined by the relation

$$\Phi\{r(\beta)\} + \left\{\frac{1}{r(\beta)} - \frac{1}{q(\beta)}\right\}\phi\{r(\beta)\}, \tag{14}$$

where the symbols Φ and ϕ mean the cumulative distribution function and the probability distribution function of a standard normal distribution, respectively. The advantage of these approximations is that they have relative error of order $O(n^{-3/2})$ in the centre of the distribution and of order $O(n^{-1})$ in the tails [8].

5. Simulation Study

The following simulation study is focused on the accuracy of the likelihood ratio test (LRT), the Wald test (W), and the score test (S), together with the proposed approximations based on the Barndorff-Nielsen (BN) and Lugannani-Rice (LR) formulas, where the likelihood root is combined with the Wald statistic (denoted as BNW and LRW, respectively), or with the score statistic (denoted as BNS and LRS, respectively). The significance of an effect of one covariate in the Cox proportional hazards model is verified under conditions of decreasing the sample size, and increasing proportion of right-censored and left-truncated data. All computations were done using the procedure PROC PHREG in the software SAS 9.3 with the Breslow approximation of the partial likelihood. Convergence was achieved for all replications under each of the model.

For each situation, 1000 independent samples were generated with the sample size $n = (100, 70, 50, 30, 20)$. Survival data consists of triplets (x_i, t_i, δ_i), $i = 1, \ldots, n$, where x_i is the value of the covariate, t_i is the observed time-to-event, and δ_i is an indicator of the survival status (event or right-censoring) for the i-th individual. Two most typically used distributions in survival analysis were considered. These are the Weibull distribution and the exponential distribution. Main characteristics such as the probability distribution function, survival and hazard function of these distributions are described in Table 1. The survival function indicates the probability that an individual survives a specified time. Together with the Gompertz distribution these follow the assumption of the proportionality for the Cox proportional hazards model. The exponential distribution is a special case of the Weibull distribution, $Wei(\lambda, \nu)$, with the shape parameter $\nu = 1$. The exponential baseline hazard function is constant unlike the Weibull one which is decreasing for $\nu < 1$ and for $\nu > 1$ is increasing. The shape parameter was chosen as $\nu = 0.5$ and $\nu = 2$; the scale parameter λ was chosen as $\lambda_1 = 1.7$, $\lambda_2 = 0.7$, and $\lambda_3 = 0.07$. The scale parameter λ determinates the variability. If λ is large, the distribution is more spread out, and, consequently, a time scale in survival analysis is long. Conversely, the distribution is more concentrated for small λ, and, thus a time scale is short, see Figure 2, where the Kaplan-Meier estimates [26] of the survival functions for the Weibull baseline hazard functions $Wei(1.7, 0.5)$ and $Wei(1.7, 2)$ are presented. For $Wei(1.7, 2)$, a time scale is six times longer. The survival functions are categorized by a dichotomous variable x. While the category $x = 1$ demonstrates better survival performance for $Wei(1.7, 0.5)$, for $Wei(1.7, 2)$ the opposite is true. Both survival functions are without censored observations and without truncation.

One covariate x was considered either as a dichotomous predictor, or as a continuous. Effects of a continuous covariate were generated from the standard normal distribution $N(0, 1)$; for a dichotomous one, effects were generated from the binomial distribution $Bi(1, 0.6)$. Times to event were generated from the Cox model assuming no effect of a covariate on survival time (the true value of the regression parameter is $\beta = 0$) and the exponential and the Weibull baseline hazard function using the inverse probability method [6], see Table 2. Right-censoring δ_i was generated randomly assuming 0, 20, 50, and 70 percent of censored observations. If $\delta_i = 0$, i.e. the i-th individual was censored, the censoring time c_i was generated in the same way as the time of event t_i together with the additional condition $c_i \leq t_i$, and (x_i, c_i, δ_i) was considered as the resulting data triplet for the i-th individual. Right-censoring due to leaving the study from other causes than an event was only considered. Left-truncation was chosen according to the time scale and it

Table 1: Weibull and exponential distribution characteristics.

Characteristic	Exponential distribution	Weibull distribution
Parameters	$\lambda > 0$	$\lambda > 0$ and $\nu > 0$
Support	$[0, \infty)$	$[0, \infty)$
Density	$f_0(t) = \lambda \exp(-\lambda t)$	$f_0(t) = \lambda \nu t^{\nu-1} \exp(-\lambda t^{\nu})$
Survival function	$S_0(t) = \exp(-\lambda t)$	$S_0(t) = \exp(-\lambda t^{\nu})$
Hazard function	$h_0(t) = \lambda$	$h_0(t) = \lambda \nu t^{\nu-1}$
Mean	λ^{-1}	$\lambda^{-(1/\nu)}\Gamma(\frac{1}{\nu} + 1)$

Table 2: Characteristics for the Cox model with the exponential and the Weibull baseline hazard function, $V \sim \mathcal{U}(0, 1)$.

Characteristics	Exponential distribution	Weibull distribution
Time-to-event	$T = -\frac{\log(V)}{\lambda \exp(x\beta)}$	$T = \left[-\frac{\log(V)}{\lambda \exp(x\beta)}\right]^{1/\nu}$
Hazard function	$h(t, x) = \lambda \exp(x\beta)$	$h(t, x) = \lambda \exp(x\beta)\nu t^{\nu-1}$

was designed to be 0, 5, 15, and 25 percent of observations were truncated. It means, the truncated time point was considered as the 0, 5-th, 15-th, and 25-th percentile of each of a time scale.

On the significance level $\alpha = 0.05$, the true null hypothesis $H_0 : \beta = 0$ was verified and the relative frequencies of rejecting the null hypothesis were investigated. Observed relative frequencies of rejecting a true H_0 were further compared with the confidence interval for a proportion of successes in 1000 binomial trials on the confidence level 0.95. The test and its p-value was considered to be conservative if a relative frequency of rejecting a true null hypothesis was smaller than 0.037, and as a liberal, when a relative frequency was greater than 0.064. Otherwise, the test and its p-value was considered as accurate. As was mentioned above, there is also possibility to assess the accuracy of tests graphically using plotting the empirical cumulative distribution function of the p-value on the vertical axis versus the theoretical one on the horizontal axis. If the p-value is accurate, the curve for the empirical cumulative distribution function of the p-value nearly coincides with the identity line. The curve for a conservative p-value is above the identity line; the curve below the identity line indicates a liberal p-value [38].

For one chosen situation, the observed relative frequencies of rejecting the true null hypothesis per 1000 simulations are presented in Table 3. Particularly, the results for the Cox model with a dichotomous covariate and with the Weibull baseline hazard function $Wei(1.7, 0.5)$ for the sample size $n = 50$ are shown. The results indicate the lowest accuracy for the score test, which was in more than half cases liberal. Consequently, the Lugannani-Rice formula and the Barndorff-Nielsen formula in combination with the score statistic were still in some cases liberal. The likelihood ratio test tended to be liberal as well. The highest accuracy was achieved for the Wald test (conservative for truncation 25% and censoring 70% only) and the Lugannani-Rice formula and the Barndorff-Nielsen formula in combination with the Wald statistic which were considered as accurate in all cases. Similar results were also obtained for other distributions and for a continuous predictor. The tests became increasingly inaccurate with decreasing sample size, as we can see in Figure 1, where the empirical cumulative distribution function of p-value versus the theoretical cumulative distribution function of $\mathcal{U}(0, 1)$ for the Cox proportional hazards model with one dichotomous covariate with the Weibull baseline hazard function $Wei(1.7, 0.5)$ are presented. Apparently, the Wald test tended to be conservative in the small-sample cases and the approximation with LR and BN improved accuracy. The behaviour of p-values of some tests are not visible because they overlap. Specifically, in panels (a) and (b), BNW with LRW, and BNS with LRS are overlapping. For large samples, the accuracy of

Table 3: Observed relative frequencies of rejecting the true null hypothesis about the regression coefficient β on the significance level $\alpha = 0.05$ for a dichotomous covariate in the Cox model with the Weibull baseline hazard function $Wei(1.7, 0.5)$ for a sample size $n = 50$ per 1000 simulations.

Truncation(%)	Censoring(%)	LRT	S	W	BNS	BNW	LRS	LRW
0	0	0.061	0.063	0.059	0.061	0.061	0.061	0.061
	20	0.062	0.069	0.062	0.064	0.064	0.064	0.064
	50	0.057	0.060	0.055	0.058	0.058	0.058	0.058
	70	0.066	0.066	0.044	0.067	0.056	0.067	0.055
5	0	0.063	0.065	0.060	0.064	0.062	0.064	0.062
	20	0.061	0.064	0.061	0.062	0.061	0.062	0.061
	50	0.060	0.064	0.057	0.062	0.060	0.062	0.060
	70	0.060	0.062	0.041	0.059	0.051	0.059	0.049
15	0	0.061	0.066	0.061	0.060	0.060	0.060	0.060
	20	0.063	0.069	0.062	0.065	0.062	0.065	0.062
	50	0.056	0.063	0.055	0.056	0.055	0.056	0.055
	70	0.064	0.066	0.044	0.065	0.054	0.065	0.050
25	0	0.065	0.063	0.062	0.063	0.063	0.063	0.063
	20	0.063	0.067	0.062	0.064	0.063	0.064	0.063
	50	0.061	0.066	0.052	0.063	0.059	0.063	0.059
	70	0.071	0.075	0.034	0.070	0.054	0.070	0.048

all considered test is correct.

Summary of relative frequencies of observed inaccurate p-values (liberal and conservative) calculated for a given covariate and distribution of baseline hazard function over all groups of censoring and truncation are demonstrated in Tables 4 and 5. It easy to see that the Wald test is mostly conservative and overestimates the true p-value, while the likelihood ratio test and the score test are mostly liberal and underestimate the true p-value. For a dichotomous covariate these tendencies are more visible. The Barndorff-Nielsen and the Lugannani-Rice approximations in combination with the Wald test (BNW, LRW) give better results and they have less inaccurate p-values than the other tests. The results for a continuous covariate indicate that the Wald test is a good choice for hypothesis testing about a scalar parameter. Note, that this test is the main one in hypothesis testing in SAS under PROC PHREG. The combinations of the likelihood root with the score statistic, LRS and BNS, do not improve the accuracy.

6. Conclusions and Discussions

The paper was inspired by the real-world application of the Cox proportional hazards model with right-censored and left-truncated data, where the effect of the size of the mitral valve prothesis on time-to-event was analysed. Truncation in a situation like this can completely change the meaning of the model in terms of the effect significance [5]. Therefore the accuracy of tests about significance of one regression coefficient was studied. The paper is focused on often used large-sample tests like the likelihood ratio test, the Wald test and the score test. The results from the large simulation study indicate that the Wald test is mostly conservative in contrast to the likelihood ratio test and the score test whose are liberal. The accuracy of tests decreases with increasing proportion of right-censored and left-truncated data and decreasing sample size. The accuracy in testing the effects of a dichotomous covariate is less than for a continuous one. The score statistic is more inaccurate than the likelihood ratio statistic.

For the improvement of the accuracy of these tests, higher order approximations of the likelihood root based on the Lugannani-Rice and the Barndorff-Nielsen formula were proposed. The simulations showed

Table 4: Summary of relative frequencies (in percentages) of observed inaccurate p-values (conservative/liberal) for a dichotomous covariate.

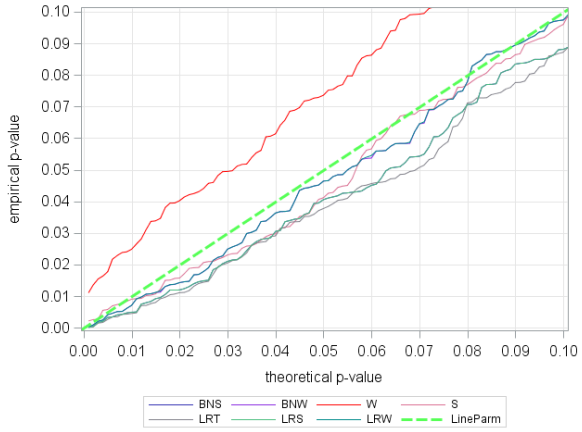
Distribution	LRT	Score	Wald	BNS	BNW	LRS	LRW
Wei(0.7,2)	0/12.5	0/13.75	18.75/1.25	0/15	0/7.5	0/15	12.5/1.25
Wei(1.7,2)	0/13.75	0/15	27.5/1.25	0/12.5	1.25/8.75	0/12.5	13.75/2.5
Wei(0.07,2)	0/21.25	0/23.75	0/23.75	0/22.5	0/12.5	0/22.5	11.25/6.25
Wei(0.7,0.5)	0/17.5	0/27.5	0/16.25	0/17.5	0/11.25	0/17.5	12.5/3.75
Wei(1.7,0.5)	0/38.75	0/56.25	17.5/5	0/45	0/23.75	0/45	10/15
Wei(0.07,0.5)	1.25/27.5	1.25/26.25	23.75/7.5	1.25/26.25	2.5/18.75	1.25/26.25	13.75/11.25
Wei(0.7,1)	0/33.75	0/45	16.25/7.5	0/38.75	0/22.5	0/38.75	12.5/16.25
Wei(1.7,1)	1.25/25	0/31.25	21.25/5	1.25/26.25	1.25/12.5	1.25/26.25	12.5/6.25
Wei(0.07,1)	0/30	0/37.5	18.75/7.5	0/35	0/16.25	0/35	10/10

Table 5: Summary of relative frequencies (in percentages) of observed inaccurate p-values (conservative/liberal) for a continuous covariate.

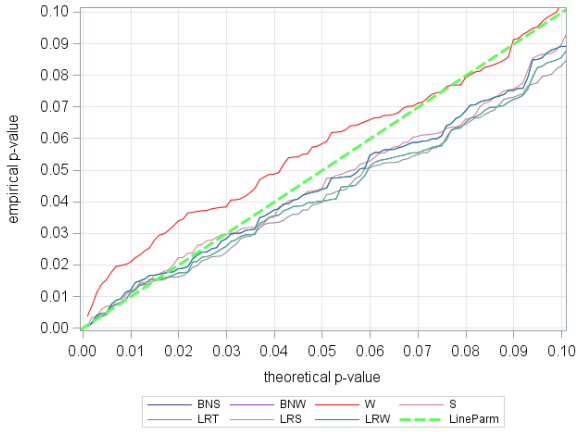
Distribution	LRT	Score	Wald	BNS	BNW	LRS	LRW
Wei(0.7,2)	0/18.75	0/7.5	7.5/2.5	0/17.5	0/12.5	0/17.5	0/10
Wei(1.7,2)	0/18.75	0/8.75	8.75/3.75	0/16.25	0/12.5	0/16.25	0/11.25
Wei(0.07,2)	0/13.75	0/7.5	7.5/1.25	0/12.5	0/10	0/12.5	0/7.5
Wei(0.7,0.5)	0/20	0/11.25	10/2.5	0/15	0/11.25	0/15	1.25/6.25
Wei(1.7,0.5)	0/27.5	2.5/10	16.25/5	0/22.5	1.25/17.5	0/22.5	1.25/13.75
Wei(0.07,0.5)	0/26.25	1.25/13.25	0/10	0/26.25	0/20	0/26.25	2.5/15
Wei(0.7,1)	0/20	1.25/13.75	8.75/0	0/17.5	0/12.5	0/17.5	0/8.75
Wei(1.7,1)	0/32.5	0/20	10/10	0/32.5	0/22.5	0/32.5	0/18.75
Wei(0.07,1)	0/15	0/7.5	11.25/0	0/13.75	0/8.75	0/13.75	1.25/5

Figure 1: Accuracy of p-values of the likelihood ratio test (LRT), the Wald test (W), the score test (S) together with the approximations based on the Lugannani-Rice (LR) and Barndorff-Nielsen (BN) formula for the Cox proportional hazards model with one dichotomous covariate for the Weibull baseline hazard function $Wei(1.7, 0.5)$, truncation 25 percent and censoring 50 percent.

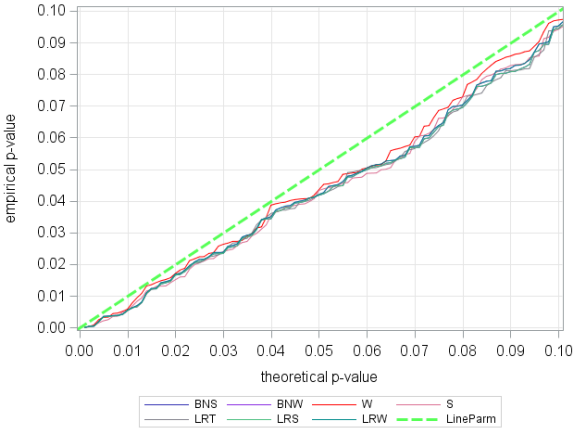
(a) sample size n=20



(b) sample size n=30



(c) sample size n=70



(d) sample size n=100

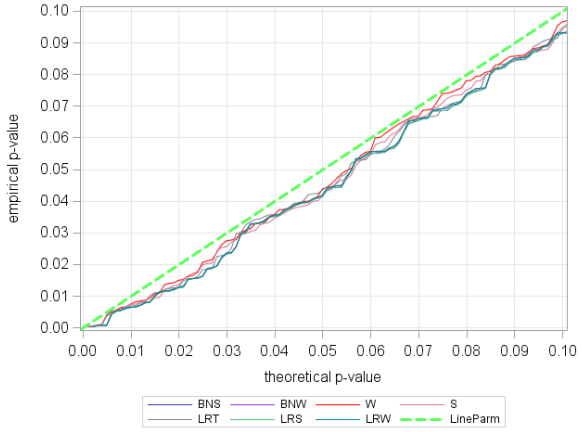
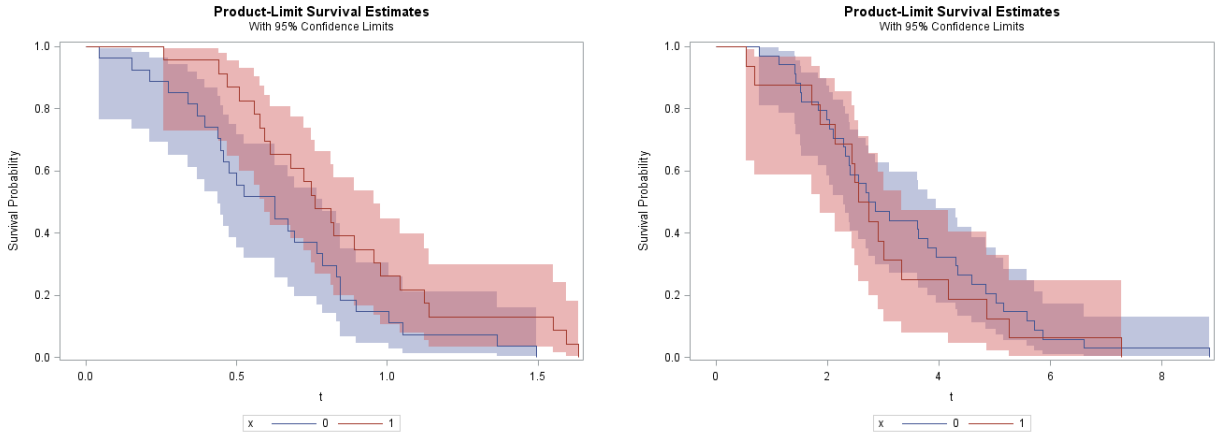


Figure 2: The Kaplan-Meier estimates of survival functions with 95% confidence interval for the Weibull baseline hazard function $Wei(1.7, 0.5)$ on the left panel, and $Wei(1.7, 2)$ on the right panel. Sample size is $n = 50$ and no censoring and truncation is considered. Survival functions are categorized by a dichotomous covariate.

(a) sample size n=50, $Wei(1.7, 0.5)$ (b) sample size n=50, $Wei(1.7, 2)$



usefulness mainly for the combination of the Lugannani-Rice formula and the Barndorff-Nielsen formula with the Wald statistic. Conversely, the combination of the likelihood root with the score statistic does not improve the accuracy.

The paper is focused on the scalar parameter of interest only, however, in practice models often include more that one covariate. In such situation, the proposed scalar approximations to each covariate of interest separately can be applied in a similar way. The nuisance vector parameter is estimated by means of the constrained maximum likelihood for a fixed parameter of interest, and, consequently, the likelihood root is computed from the profile likelihood. When inference for a vector parameter are of interest, higher order approximations based on Bartlett correction of the likelihood ratio statistic or Skovgaard's statistics may be used [8]. The extension to more than one parameter is left for further research.

References

[1] P. D. Allison, Survival Analysis Using the Sas System (A Practical Guide), SAS Institute Incorporated, 1995.
[2] P. Austin, Generating survival times to simulate Cox proportional hazards models with time-varying covariates, Statistics in Medicine 31 (2011) 3946–3958.
[3] O. Barndorff-Nielsen, D.R. Cox, Edgeworth and saddle-point approximations with statistical applications, Journal of the Royal Statistical Society (Series B) 41 3 (1979) 279–312.
[4] S. Bělašková, E. Fišerová, Detection of Influential Factors on Unemployment Duration of Tomáš Baťa University Graduates by the Hazard Model, Conference Proceedings of the 32nd International Conference Mathematical Methods in Economics (2014) 37–42 ISBN 978-80-244-4209-9.
[5] S. Bělašková, E. Fišerová, Study of bootstrap estimates in Cox regression model with delayed entry, Acta Universitatis Palackianae Olomucensis, Facultas Rerum Naturalium, Mathematica 52 (2013) 21–30.
[6] R. Bender, T. Augustin, M. Blettner, Generating survival times to simulate Cox proportional hazards model, Statistics in Medicine 24 (2005) 1713–1723.
[7] R.L. Berger, D.D. Boos, P values maximized over a confidence set for nuisance parameter, Journal of the American Statistical Association 89 (1994) 1012–1016.
[8] A.R. Brazzale, A.C. Davison, N. Reid, Applied Asymptotics: Case Studies in Small-Sample Statistics, Cambridge: Cambridge University Press 2007.
[9] N.E. Breslow, Discussion of Professor Cox's Paper, Journal of the Royal Statistical Society (Series B) 34 (1972) 216–217.
[10] N.E. Breslow, Covariance Analysis of Censored Survival Data, Biometrics 30 (1974) 89–99.
[11] N.C. Cary, SAS Institute Inc. (User's Guide), SAS Institute Inc. SAS/STAT 9.2 2008.

[12] T.K. Chandra, S.N. Joshi, Comparison of likelihood ratio, Rao's and Wald's tests and a conjecture of C.R.Rao, Sankhya A, 45 (1983) 226–246.
[13] D. Collett, Modeling Survival Data in Medical Research, London: Chapman & Hall 1994.
[14] D. Commenges, L. Letenneur, P. Joly, A. Alioum, J.F. Datigues, Modeling age-specific risk:application to dementia, Statistics in Medicine 17 (1998) 1973–1988.
[15] D.R. Cox, Regression Models and Life-Tables, Journal of the Royal Statistical Society (Series B) 34(2) (1972) 187–220.
[16] D.R. Cox, D.V Hinkey, Theoretical Statistics, London: Chapman & Hall (1974).
[17] D.R. Cox, D. Oakes, Analysis of Survival Data, London: Chapman & Hall (1984).
[18] B. Efron, R.J. Tibshirani, An introduction to the bootstrap, New York: Chapman & Hall (1993).
[19] E. Fišerová, M. Chvosteková, S. Bělašková, M. Bumbálek, Z. Joska, Survival Analysis of Factors Influencing Cyclic Fatigue of Nickel-Titanium Endodontic Instruments, Advanced in Materials Science and Engineering 189703 (2015) 1–6.
[20] D.A.S. Fraser, N. Reid, J. Wu, A simple general formula for tail probabilities for frequentist and Bayesian inference, Biometrica (1999).
[21] M.H. Gail, B. Graubard, D.F. Williamson, K.M. Flegal, Comment on Choice of time scale and its effect on significance of predictors in longitudinal studies, Statistics in Medicine 28(8) (2009) 1315–1317.
[22] J.C. Gardiner, Survival Analysis: Overview of parametric, nonparametric and semiparametric approaches and new developments, SAS Global forum (2010) 252–2010.
[23] P. Hu, A.A. Tsiatis, M. Davidian, Estimating the parameters in the Cox model when covariate variables are measured with error, Biometrics 54 (1998) 1407–1419.
[24] D.D. Ingram, D.M. Makuc, Statistical issues in analysing the NHANES I Epidemiologic Follow-up Study, Journal of Applied Probability and Statistics (2012) 87–108.
[25] M.A. Jamali, H. Voghouei,N.G. Nor Mohd, Information technology and survival of SMEs: an empirical study on Malaysian manufacturing sector, Information Technology and Management 16 2 (2015) 79–95.
[26] E.L. Kaplan, P. Meier, Non-parametric Estimation from Incomplete Observations, Journal of the American Statistical Association 53 (1958) 457–481.
[27] J.P. Klein, M.L. Moeschberger, Survival Analysis (Techniques for Censored and Truncated Data), Springer New York 1997.
[28] D.G. Kleinbaum, M. Klein, Survival Analysis (A Self-Learning Text, Second Edition), New York: Springer-Verlag 2005.
[29] E.L. Korn, B.I. Graubard, D. Midhune, Time-to-event analysis of longitudinal follow-up of a survey (choice of time scale), American Journal of Epidemiology 145(1) (1997) 72–80.
[30] W.R. Lane, S.W. Looney, J.W. Wansley, An application of the cox proportional hazards model to bank failure, Journal of Banking & Finance 10 4 (1986) 511–531.
[31] E.T. Lee, O.T. Go, Survival analysis in public health research, Annual Review of Public Health 18 (1997) 105–134.
[32] E.T. Lee, W.J. Wenyu, Statistical methods for survival data analysis–3rd ed., Wiley 2003.
[33] R. Lugannani, S. Rice, Saddle point approximation for the distribution of the sum of independent random variables, Advances in Applied Probability 12 (1980) 475–490.
[34] H.W. Peers, Likelihood ratio and associated test criteria, Biometrica 58, 577–587.
[35] T.M. Therneau, A Package for Survival Analysis in S (Technical Report 53), Section of Biostatistics, Mayo Clinic (1994).
[36] A.C. Thiébaut, J. Bénichou, Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study, Statistics in Medicine, 23 (2004) 3803–3820.
[37] J. Volaufova, L. LaMotte, Comparison of approximate tests of fixed effects in linear repeated measures design models with covariates, Tatra Mountains 39 (2008) 17–25.
[38] J. Volaufova, Accuracy of p-values of approximate tests in testing for equality of means under unequal variances, Mathematica Slovaca 59 (2009) 679–692.
[39] C.Y. Wang, L. Hsu, Z.D. Feng, L.R. Prentice, Regression calibration in failure time regression, Biometrics 53 (1997) 131–145.
[40] Y. Yi, X. Wang, Comparison of Wald, score and likelihood ratio tests for response adaptive designs, Journal of the Statistical Theory and Applications 10(4) (2011) 553–569.