STUDIA OECONOMICA POSNANIENSIA 2014, vol. 2, no. 2 (263)

Jadwiga Borucka
Warsaw School of Economics, Institute of Statistics and Demography,
Event History and Multilevel Analysis Unit
jadwiga.borucka@gmail.com

METHODS FOR HANDLING TIED EVENTS IN THE COX PROPORTIONAL HAZARD MODEL

Abstract: The Cox proportional hazard model is one of the most common methods used in time to event data analysis. The model is based on several restrictive assumptions one of which concerns tied events, i.e. events with exactly the same survival time. If time were measured in a perfectly continuous scale, such cases would never occur. In real applications time is usually measured in a discrete manner which results in the existence of ties in most survival data. However, if this assumption is violated, it should not necessarily prevent analysis by using the Cox model. The current paper presents and compares five ways proposed for handling tied events. On the basis of the calculations performed it can be stated that exact expression and the discrete model give the best results in terms of fit statistics; however, they are the most time-consuming. Efron and Breslow approximations are much faster but result in worse model fit. In the case analysed the Efron method seems to be the best choice, taking into account differences in parameters estimates, fit statistics and calculation time. What is more, a simple method based on subtracting a tiny random value from each tied survival time performed surprisingly well; both in terms of parameter estimates compared with the exact expression, as well as fit statistics. In general, in the case of large datasets and/or a large number of ties, if estimation precision is not as important as estimation time, Breslow or – more preferably – Efron approximations might be used. However, if time is not limited, one should consider choosing an exact method or discrete model that can provide better fit statistics and more efficient parameter estimates.

Keywords: Cox model, tied events, applied survival analysis, partial likelihood function.

JEL Classification: C13, C14.

METODY ESTYMACJI MODELU PROPORCJONALNYCH HAZARDÓW COXA W WYPADKU WYSTĘPOWANIA ZDARZEŃ POWIĄZANYCH

Streszczenie: Model proporcjonalnych hazardów Coxa jest jedną z najczęściej wykorzystywanych metod w analizie czasu przeżycia. Jest oparty na kilku restrykcyjnych założeniach; jedno z nich dotyczy występowania tzw. zdarzeń powiązanych, tzn. zdarzeń zaobserwowanych dokładnie w tym samym momencie. Gdyby czas był mierzony w sposób ciągły, taka sytuacja nie miałaby miejsca. W zastosowaniach praktycznych jednakże czas jest zazwyczaj mierzony w sposób dyskretny, co skutkuje występowaniem zdarzeń powiązanych w większości zbiorów wykorzystywanych w analizie czasu przeżycia. Niemniej jednak, odchylenie od tego założenia nie musi stanowić przeszkody w stosowaniu modelu Coxa. Niniejszy artykuł prezentuje pięć sposobów proponowanych w literaturze dla zbiorów zawierających zdarzenia powiązane. Na podstawie przeprowadzonych obliczeń można stwierdzić, że metoda *exact* oraz model dyskretny dają najlepsze rezultaty pod względem statystyk dopasowania, jednakże są najbardziej czasochłonne. Przybliżenia Breslowa i Efrona pozwalają otrzymać oszacowania znacznie szybciej, ale skutkują gorszą jakością dopasowania modelu. W rozważanym przykładzie przybliżenie Efrona jest najlepszym wyborem, biorąc pod uwagę różnice w oszacowaniach parametrów, statystyki dopasowania modelu oraz czas obliczeń. Ponadto prosta metoda oparta na odjęciu od każdej wartości zmiennej czasowej pośród zdarzeń powiązanych niewielkiej wartości wylosowanej z rozkładu jednostajnego, daje zaskakująco dobre wyniki zarówno pod względem oszacowań parametrów, jak i statystyk dopasowania modelu. Ogólnie rzecz biorąc, w przypadku dużych zbiorów i/lub dużej liczby zdarzeń powiązanych, jeśli precyzja estymacji nie jest priorytetem, a ważniejszy jest czas obliczeń, przybliżenia Breslowa lub Efrona mogą być wykorzystane. Jednakże jeśli czas nie jest ograniczony, warto rozważyć wybór metody *exact* lub modelu dyskretnego, które umożliwiają otrzymanie lepszych statystyk dopasowania modelu oraz wyższej efektywności oszacowań parametrów.

Słowa kluczowe: model Coxa, zdarzenia powiązane, analiza czasu przeżycia, funkcja częściowej wiarygodności.

Introduction

The Cox proportional hazard model is one of the most common methods used in time to event data analysis. The idea of the model is to define hazard level as a dependent variable which is being explained by the time-related component (so called baseline hazard) and covariate-related component. The model is defined as follows:

$$h(t, x, \beta) = h_0(t) \exp(\beta x),$$

where:

$h(t, x, \beta)$ – hazard function that depends on timepoint t and vector of covariates x,

$h_0(t)$ – baseline hazard function that depends on time only,
$\exp(\beta x)$ – covariate-related component,
β – vector of parameter estimates.

The model is based on several restrictive assumptions one of which concerns tied events, i.e. events with exactly the same survival time. This assumption is directly related to the method of Cox model estimation. The parameter estimates are obtained through the use of the maximum likelihood method, as suggested by Cox [1972]. Assuming a simple model for one event with right censoring, including a single covariate and using the following relationship between the probability density function of the time variable, hazard function and survival function:

$$f(t, x, \beta) = h(t, x, \beta) \cdot S(t, x, \beta)$$

the likelihood function for Cox model can be derived as follows [Hosmer & Lemeshow 1999]:

$$l(\beta) = \prod_{i=1}^{n} \left\{ \left[h(t_i, x_i, \beta) \right]^{c_i} \left[S(t_i, x_i, \beta) \right]^{1-c_i} \right\},$$

where:

$l(\beta)$ – likelihood function depending on parameter β,
$i = 1, 2, \ldots, n$ – observations ordered by time (actual or censored),
$h(t_i, x_i, \beta)$ – hazard function at timepoint t_i for subject with covariate value x_i,
$S(t_i, x_i, \beta)$ – survival function at timepoint t_i for subject with covariate value x_i,
c_i – censoring indicator (equal to 1 if subject experienced the event and 0 if subject is censored).

After taking the logarithm and doing certain transformations the following expression is obtained:

$$L(\beta) = \sum_{i=1}^{n} \left\{ c_i \ln\left[h_0(t_i) \right] + c_i x_i \beta + e^{x_i \beta} \ln\left[S_0(t_i) \right] \right\}.$$

The full likelihood function as defined above requires maximization with respect to the parameter β as well as to the baseline hazard function which

is unspecified. Cox [1972] suggests using an alternative expression, called by him the partial likelihood function, which depends on the parameter β only. He argues that the estimation of parameter β obtained through the use of the function proposed by him should have the same properties as one that would result from the full likelihood function. This thesis is proved in Andersen et al. [1993] as well as Fleming and Harrington [1991]. The approach presented by Cox leads to the following formula for the partial likelihood function:

$$l_p(\beta) = \prod_{i=1}^{n} \left[\frac{e^{x_i \beta}}{\sum_{j \in R(t_i)} e^{x_j \beta}} \right]^{c_i},$$

where $R(t_i)$ – risk set at timepoint t_i, which can also be rewritten, taking into account only non-censored observations, as follows:

$$l_p(\beta) = \prod_{i=1}^{m} \frac{e^{x_{i^*} \beta}}{\sum_{j \in R(t_{i^*})} e^{x_j \beta}},$$

where:

$l_p(\beta)$ – partial likelihood function depending on parameter β,
$i = 1, 2, \ldots, n$ – non-censored observations ordered by time (actual),
t_{i^*} – survival time of i^{th} subject who experienced the event.

The above formula results in the following form for the logarithm of partial likelihood function:

$$L_p(\beta) = \sum_{i=1}^{m} \left\{ x_{i^*} \beta - \ln \left[\sum_{j \in R(t_{i^*})} e^{x_j \beta} \right] \right\}$$

and – after differentiating with respect to β [Hosmer & Lemeshow 1999]:

$$\frac{\partial L_p(\beta)}{\partial \beta} = \sum_{i=1}^{m} \left\{ x_{i^*} - \frac{\sum_{j \in R(t_{i^*})} x_j e^{x_j \beta}}{\sum_{j \in R(t_{i^*})} e^{x_j \beta}} \right\} = \sum_{i=1}^{m} \left\{ x_{i^*} - \sum_{j \in R(t_{i^*})} w_{ij}(\beta) x_j \right\} =$$

$$= \sum_{i=1}^{m} \left\{ x_{i^*} - \overline{x}_{w_i} \right\},$$

where:

$$w_{ij}(\beta) = \frac{e^{x_j\beta}}{\sum_{l \in R(t_{i^*})} e^{x_l\beta}} \quad \text{and} \quad \overline{x}_{w_i} = \sum_{j \in R(t_{i^*})} w_{ij}(\beta) x_j.$$

On the basis of the equation above it can be seen that the order of events does matter: each subject who experiences the event has their own contribution to the partial likelihood function which includes summing up some expressions for all the subjects who are at risk at the moment at which the event for this particular subject is being observed. Let us say there are two subjects A and B experiencing an event at exactly the same time. In such a situation it is not clear whether subject A should be considered as being at risk while subject B is experiencing the event and vice versa. Thus, the partial likelihood function defined as above assumes that there are no tied events among the observations. If time were measured in a perfectly continuous scale, such cases would never occur and a proper ordering of events would not be a problem. In fact, in most applications, time is being measured in a discrete manner which results in the existence of ties in most survival data. However, even if the assumption of lack of tied events is violated, it does not exclude the Cox model as a potential tool in an analysis of time to event data, although deviation from the assumption should not be neglected. So far there have been several methods developed to handle tied events in proportional hazard models. The next section of the current paper presents a theoretical background for four of them: the Breslow approximation, the Efron approximation, Kalbfleisch and Prentice exact expression, as well as the discrete model. Additionally, a simple alternative method based on subtracting tiny random value from each tied survival time is described. In the subsequent part of the article, these methods are implemented in practice and compared in terms of differences in parameter estimates, efficiency of parameters estimates, fit statistics and computational time.

1. Handling tied events – theoretical background

In order to take into account tied events in Cox model estimation it is necessary to adjust the partial likelihood function appropriately. As Kalbfleisch and Prentice [2002] argue, the most natural way to do this is to consider all possible orders of event occurrences for subjects having the same survival

time. Their idea can be described as follows by assuming there are m distinct observed survival times ordered as follows: $t_1 < \ldots < t_m$ and that d_i events happen at t_i, where $i = 1, 2, \ldots, m$. Furthermore, Kalbfleisch and Prentice define $D(t_i) = \{i_1, \ldots, i_{d_i}\}$ as the set of labels of failing observations at t_i. Taking Q_i as the set of $d_i!$ permutations for d_i events observed at t_i, P is one element of Q_i and is defined as $P = (p_1, \ldots, p_{d_i})$ and finally $R(t_i, P, k) = R(t_i) - \{p_1, \ldots, p_{k-1}\}$. Then, each subject with an observed survival time equal to t_i has a contribution to the partial likelihood function equal to:

$$\frac{1}{d_i!} e^{x_{i+}\beta} \sum_{P \in Q_i} \prod_{k=1}^{d_i} \left\{ \sum_{l \in R(t_i, P, k)} e^{x_l \beta} \right\}^{-1},$$

where $x_{i+} = \sum_{j=1}^{d_i} x_j$, which results in the following formula for the partial likelihood function:

$$l(\beta) = \prod_{i=1}^{m} \frac{e^{x_{i+}\beta}}{\sum_{P \in Q_i} \prod_{k=1}^{d_i} \left[\sum_{j \in R(t_i, P, k)} e^{x_j \beta} \right]}.$$

Maximization of the function defined as above might be time-consuming, especially in the case of a large number of ties. This fact led to the derivation of some approximate expressions for the partial likelihood function. One of them is presented by Breslow [1975] who suggests summing up covariate-related components for all subjects experiencing the event at a given timepoint t_i and raising the result to a power equal to the number of events tied at t_i. The partial likelihood function that uses this approach is defined as follows:

$$l(\beta) = \prod_{i=1}^{m} \frac{e^{x_{i+}\beta}}{\left[\sum_{j \in R(t_i)} e^{x_j \beta} \right]^{d_i}}.$$

However, if the number of tied events for any timepoint is relatively large, this method might not give a good approximation of the partial likelihood function as defined by Kalbfleisch and Prentice. An alternative suggestion comes

from Efron [1977]. According to his idea, the partial likelihood function can be approximated as follows:

$$l(\beta)=\prod_{i=1}^{m}\frac{e^{x_{i+}\beta}}{\prod_{k=1}^{d_i}\left(\sum_{j\in R(t_i)}e^{x_j\beta}-\frac{k-1}{d_i}\sum_{j\in D(t_i)}e^{x_j\beta}\right)}.$$

As Kalbfleisch and Prentice point out, obtaining parameter estimates on the basis of the above defined approximation is not particularly complicated; however, the estimators resulting from Efron and Breslow approximations might be biased. What is more, estimator of the variance of parameter $\hat{\beta}$ is not consistent [Kalbfleisch & Prentice 2002]. On the basis of numerous calculations it has been shown that the Breslow method causes severe bias for datasets with large fraction of d_i/n_i, where n_i is the number of subjects at risk at t_i; while Efron approximation and exact expression still perform well, even if number of ties is high. If the number of tied events is very small, all three methods give very similar results. For datasets with no ties, all methods lead to exactly the same results.

As an alternative for approximations of the partial likelihood function, applicable especially for datasets with a large number of ties (which might suggest the fact that tied events do not result from insufficient precision in time measurement but rather from the discrete character of the time variable), one can consider choosing the approach suggested by Cox [1972], i.e. using a discrete logistic model defined as follows:

$$\frac{h(t,x,\beta)}{1-h(t,x,\beta)}=\frac{h_0(t)}{1-h_0(t)}e^{x\beta}.$$

Kalbfleisch and Prentice provide a generalization of the partial likelihood function applicable to the model defined above. It enables a calculation of the probability that, given the risk set at timepoint t_i and the number of events d_i occurring at this timepoint, the events will be experienced by exactly these subjects who have an observed survival time equal to t_i. The conditional probability is given by [Kalbfleisch & Prentice 2002]:

$$\prod_{i=1}^{m}\frac{e^{x_{i+}\beta}}{\sum_{j\in R_{d_i(t_i)}}e^{x_{i+}\beta}},$$

where $R_{d_i(t_i)}$ constitutes the set of all possible subsets including exactly d_i distinct subjects selected out of units who are at risk at t_i, denoted as $R(t_j)$.

The methods described above are usually mentioned in the literature as far as the handling of tied events in the Cox model is concerned. As a simple alternative, there is another way proposed in the current paper. The method itself is suggested by Hosmer and Lemeshow [1999] as the way to break ties for the purpose of non-parametric survival function estimation. The idea is to subtract a tiny random value from each tied survival time. In this way, tied events will be uniquely ordered with respect to each other but their position in relation to all remaining observations remains unchanged. Hosmer and Lemeshow argue that this solution has no effect when estimating the survival function as the estimate for the last tied event at t_i is exactly the same as if it were calculated for all ties at t_i considered simultaneously. They do not analyse however the utility of this method in terms of breaking ties in the Cox model. In the current paper, the idea of this method is being applied for a Cox model where tied events cannot be neglected due to partial likelihood function requirements. In order to break ties, a random value from the uniform distribution defined at the interval [0, 0.001] is subtracted from each tied observed survival time. In this case, it does not matter which version of the partial likelihood function is being used (Efron, Breslow or exact expression) as all of them lead to the same results if there are no ties among the observations.

2. Methods of tied events handling – application and results

As has been mentioned, the main aim of the present paper is to compare five methods described above in terms of parameter estimates, efficiency of estimators, fit statistics, as well as computational time. In order to do this, the same Cox proportional hazard model was estimated through the use of each of these methods. Additionally, for each method the estimation was repeated 10 times and the average computational time per method was calculated. The dataset used for analysis contained artificially generated data including the following: time variable, censoring indicator and 7 covariates that are considered as potential explanatory variables in time to event analysis. The dataset contains 6500 observations, out of which 1700 are censored. The number of ties is relatively high: there are 31 distinct observed survival times, their counts ranging from 100 to 500. Due to the fact that the present paper is focused on the methodology rather than on empirical results, using artificial data is justified. Parameter estimates are not supposed to be interpreted as the main

aim is to compare the five methods of handling tied events with regard to the estimator efficiency and fit statistics, thus, the variables in the dataset are not named directly but are referred to as time variable, censoring indicator and covariates (or explanatory variables) numbered from 1 to 7. All calculations were performed using the SAS® Base 9.3 module. The SAS code that enabled the model estimation using all of these methods and to modify the time variable by subtracting a random value from the uniform distribution is included in the Appendix. As far as the method based on subtracting a tiny random value is concerned, the Breslow version of partial likelihood function was used as in the case of a lack of tied events all approximations lead to the same results; however, the algorithm for Breslow approximation embedded in the PHREG procedure is the fastest out of all four methods.

Table 1 presents parameter estimates with *p*-values and Table 2 standard errors of parameter estimates obtained through the use of each of five methods of tied events handling.

There are some visible differences between parameter estimates obtained through the use of different methods. Efron approximation gives results that are very close to those generated by exact expression. Estimates from Breslow

Table 1. Parameter estimates for the Cox model

Covariate	Exact	Efron	Breslow	Discrete	Random value
Covariate 1	–0.0250	–0.0250	–0.0230	–0.0248	–0.0250
Covariate 2	0.0171	0.0171	0.0198	0.0234	0.0170
Covariate 3	–0.1438	–0.1438	–0.1394	–0.1573	–0.1437
Covariate 4	0.0811	0.0812	0.0915	0.0915	0.0812
Covariate 5	0.1510	0.1510	0.1479	0.1707	0.1512
Covariate 6	–0.2642	–0.2650	–0.2636	–0.2700	–0.2648
Covariate 7	3.8732	3.8692	3.6705	4.0210	3.8672
			***p*-value**		
Covariate 1	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Covariate 2	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Covariate 3	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Covariate 4	0.101	0.1004	0.0617	0.0791	0.0804
Covariate 5	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Covariate 6	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Covariate 7	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001

Source: Own calculation.

Table 2. Standard errors of parameter estimates for the Cox model

Covariate	Exact	Efron	Breslow	Discrete	Random value
Covariate 1	0.0016	0.0016	0.0016	0.0017	0.0016
Covariate 2	0.0042	0.0041	0.0041	0.0044	0.0041
Covariate 3	0.0075	0.0075	0.0074	0.0078	0.0075
Covariate 4	0.0495	0.0495	0.0490	0.0521	0.0495
Covariate 5	0.0094	0.0094	0.0092	0.0098	0.0094
Covariate 6	0.0552	0.0552	0.0547	0.0570	0.0552
Covariate 7	0.1410	0.1409	0.1389	0.1475	0.1408

Source: Own calculation.

approximation seem to give underestimations for five out of seven variables, as compared with exact expression or the Efron method. This observation is in accordance with the conclusions drawn by Hertz-Picciotto and Rockhill [1977] who also indicate a tendency for Breslow approximation to underestimate parameters in the Cox model. The discrete model tends to give higher absolute values for parameter estimates (with the exception of covariate 7). In this case, however, it should be noted that the discrete model is based on a different likelihood function than the Cox proportional hazard model which uses a partial likelihood function (with potential modification such as exact expression, the Breslow or Efron approximation), thus the parameters obtained from these models do not have exactly the same interpretation [Kalbfleisch & Prentice 2002]. The results obtained through the use of random value methods do not represent any systematic pattern, as some estimates are higher than for exact expression and some are lower; however, parameter estimates are very close to those coming from the Efron or exact method. What is interesting, is the fact that differences between the parameters obtained through the use of exact expression and the random value method are in most cases lower than the differences between parameters coming from the exact method, Breslow approximation or the discrete model, which might indicate that simply subtracting a tiny random value might even give better results than some more formal methods of handling tied events. As far as standard errors are concerned, the results obtained do not differ between the five methods to a great extent, only the discrete method shows slightly higher errors, but these differences are not very high. What is worth being mentioned is that the *p*-value for covariate 3 differs between all five methods to the extent that could even lead to different conclusions concerning the statistical significance of

the covariates, depending on the significance level. Assuming alpha = 0.05 all methods indicate a lack of significance, however taking alpha = 0.1 we would have to change this decision on the basis of Breslow, discrete and random value methods. The results obtained through the use of Breslow approximation gave the strongest reason for a rejection of the null hypothesis, while those from the discrete model and random value method are not that obvious.

Additionally, standardized measures of variability, as suggested by Nardi and Schemper [2003] and defined as:

$$SV = \frac{\hat{\sigma}_{\beta}}{|\hat{\beta}|}$$

were obtained. They can be used to assess efficiency of parameter estimators. Their values are presented in Table 3.

Table 3. Standardized measure of variability

Covariate	Exact	Efron	Breslow	Discrete	Random value
Covariate 1	0.0657	0.0657	0.0705	0.0685	0.0657
Covariate 2	0.2427	0.2420	0.2056	0.1881	0.2432
Covariate 3	0.0522	0.0522	0.0531	0.0496	0.0522
Covariate 4	0.6098	0.6088	0.5352	0.5694	0.6090
Covariate 5	0.0622	0.0621	0.0623	0.0576	0.0621
Covariate 6	0.2089	0.2081	0.2074	0.2112	0.2082
Covariate 7	0.0364	0.0364	0.0378	0.0367	0.0364

Source: Own calculation.

On the basis of the standardized measure of variability, it is hard to interchangeably indicate the best method. What is worth being mentioned is that the Breslow method gave the highest value for the coefficient the most often: for four out of seven variables; however, variability is also the lowest for two variables when the Breslow approximation is applied. The lowest SV values are displayed by the discrete model in three variables; and in the random value method, exact expression and Efron approximation – in two out of seven covariates.

Additionally, fit statistics including doubled negative logarithm of likelihood function as well as the information criteria AIC and SBC were calculated – see Table 4.

Table 4. Fit statistics

Statistics	Exact	Efron	Breslow	Discrete	Random value
-2 LOG L	32 187.43	72 707.39	73 424.65	32 098.36	72 708.11
AIC	32 201.43	72 721.39	73 438.65	32 112.36	72 722.11
SBC	32 246.77	72 766.72	73 483.99	32 157.70	72 767.45

Source: Own calculations.

The discrete model guarantees the best values for fit statistics; however, the difference between the discrete model and exact expression is very slight.[1] Breslow and Efron approximations, as well as the random value method, result in much higher values for fit statistics, even more than twice as much as those coming from the exact or discrete models. The worst results come from the Breslow method. What is worth paying attention to is that the random value method enables not only better results to be obtained as compared with the Breslow method, but also values that are very close to those coming from Efron approximation.

Finally, the computational time that was needed to obtain the parameter estimates through the use of each of the five methods was compared. Estimation was performed by using the SAS PHREG procedure (Table 5; details concerning the SAS code that was used is included in the Appendix).

Table 5. Computational time (SAS PHREG procedure)

Time	Exact	Efron	Breslow	Discrete	Random value
Real Time (sec.)	2.984	0.099	0.075	4.191	0.101
CPU Time (sec.)	2.980	0.092	0.068	4.181	0.085

Source: Own calculations.

As was expected, more complicated methods such as exact expression or the discrete model were the most time-consuming. It took the most time to estimate the discrete model. Exact expression was faster than the discrete model but was still considerably slower than Breslow and Efron approximations. Estimation of the model with a partial likelihood based on Breslow approximation was the fastest, the method subtracting random value from

[1] It should be noted that comparisons of statistical fit between the discrete model and other methods that use a partial likelihood function should be performed with caution, as other models are based on different likelihood functions.

tied event times was slightly slower, as was the Efron approximation; however, these differences were very slight. In general, the discrete model and exact expression provide the best fit statistics but the estimation time needed to use these methods is visibly higher as compared with other methods. The exact values of estimation time depend obviously on the computer parameters; however, using the same equipment to estimate all the models enabled such comparisons to be performed. It should also be taken into account that the dataset used in the calculations had only 6500 records and the model contains 7 covariates, thus even the slowest method allows results to be obtained in a very short period of time. For more complex analysis designs and/or larger datasets, and/or a larger numbers of tied events, these differences might be even more visible and as such might constitute a more important factor when choosing the best methods for a given analysis. In the scientific field, precision of estimation is usually more important than the time that is needed to get the results; however, as far as practical applications are concerned, computational time might also need to be taken into account.

Conclusions

This simple empirical study showed that in the case of a relatively high number of ties, results obtained through the use of different methods lead to visibly different results. Efron approximation seems to give results that are the closest to the ones obtained through the use of exact expression as derived by Kalbfleisch and Prentice. On the other hand, a simple method based on subtracting tiny random value from each tied survival time provides results that do not differ from the exact results to a great extent. A comparison of standard errors and of parameter estimates indicates that the discrete model might be performing a little worse as compared with other methods. It would be hard, however, to choose the best method on the basis of standard errors as well as standardized measures of variability. In terms of fit statistics, exact expression and the discrete model are superior when compared with the others. Surprisingly, the random value method leads to a better fit than Breslow and is comparable to Efron approximation; however, these differences are not very great. When it comes to computational time, more sophisticated methods such as exact expression and the discrete model require substantially longer period of time to obtain estimates. Breslow and Efron approximations as well as the random value method took more or less the same time to perform calculations, which was visibly shorter when compared to exact expression. Considering all the

statistics analysed, in this instance the Efron method seems to be the best choice: fit statistics are visibly worse than for the exact or discrete methods; however, efficiency and the values of parameter estimates are nearly equal to the results of obtained through the use of the exact method, while the estimation time for the Efron method is remarkably shorter than for the model with partial likelihood based on the exact expression. What is worth being mentioned is that the simple random value method seems to be an attractive alternative here, taking into account the fact that parameter estimates were very close to those obtained through the use of exact expression, fit statistics were better than in Breslow approximation and almost equal to the ones obtained through the use of the Efron method, and the very short period of time needed to perform calculations. In general, in the case of large datasets and/or a large number of ties, if estimation precision is not extremely important but the estimation time is, the Breslow method might be used as it guarantees a relatively short calculation time. Efron approximation, which requires only a little more time to obtain results and provides a visibly better fit as well as results that are much closer to the ones coming from exact expression, might be an even better choice. If time is not that limited and estimation precision is a more important factor – which should usually be the case in scientific research – the exact method would be the most desirable one as it provides a much better fit statistics and higher efficiency of parameter estimates. Exact expression could also be considered as the best choice due to a construction that is based on the assumption that every ordering of tied events can happen with exactly the same probability, which is reasonable and 'the most safe' approach in the case of ties. Apart from the four methods that are usually described in the literature in the context of ties in the Cox model, a simple method based on subtracting a tiny random value from each tied survival time seems to be an attractive alternative. This method performs surprisingly well when compared to more formal ways of handling tied events, taking far less time to perform calculations, thus it might be considered as well, especially for certain preliminary analyses that do not require a strong theoretical foundation but where results are expected quite quickly.

References

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

Breslow, N., 1975, *Covariance Analysis of Censored Survival Data*, Biometrics, vol. 30, no. 1, s. 89–99.

Cox, D.R., 1972, *Regression Models and Life-Tables,* Journal of the Royal Statistical Society. Series B (Methodological), vol. 34, no. 2, s. 187–220.

Efron, B., 1977, *The Efficiency of Cox's Likelihood Function for Censored Data*, Journal of the American Statistical Association, vol. 72, no. 359, s. 557–565.

Fleming, T.R., Harrington, D.P., 1991, *Counting Process and Survival Analysis*, John Wiley & Sons, New York.

Hertz-Picciotto, I., Rockhill, B., 1977, *Validity and Efficiency of Approximation Methods for Tied Survival Times in Cox Regression*, Biometrics, vol. 53, no. 3, s. 1151–1156.

Hosmer, D.W., Lemeshow, S., 1999, *Applied Survival Analysis. Regression Modelling of Time to Event Data*, John Wiley & Sons, New York.

Kalbfleisch, J.D., Prentice, R.L., 2002, *The Statistical Analysis of Failure Time Data*, John Wiley & Sons, New York.

Nardi, A., Schemper, M., 2003, *Comparing Cox and Parametric Models in Clinical Studies*, Statistics in Medicine, vol. 22, s. 3597–3610, DOI: 10.1002/sim.1592.

Appendix

```
/*STEP 1
Modify time variable for tied events by subtracting small value from
uniform distribution [0, 1000]*/
proc sql;
        create table dist as select time, count(*) as count
        from Tied_events where censor = 1 group by time;
        create table Tied_events as select a.*, b.count
        from Tied_events as a left join dist as b
        on a.time = b.time;
quit;

data Tied_events;
        set Tied_events;
        if count>1 and censor = 1 then time_mod = time-ranuni(1234)/1000;
        else time_mod = time;
run;

/*STEP 2
Estimation*/

 /*Exact expression: option ties = exact*/
ods output ParameterEstimates = est_ex FitStatistics = fit_ex;
```

```
proc phreg data= Tied_events;
       model Time*Censor(0) = Age Ind1 Ind2 Ind3 Ind4 Ind5 Ind6
                             /ties = exact;
run;
ods output close;

 /*Breslow approximation: option ties = Breslow*/
ods output ParameterEstimates = est_br FitStatistics = fit_br;
proc phreg data= Tied_events;
       model Time*Censor(0) = Age Ind1 Ind2 Ind3 Ind4 Ind5 Ind6
                            /ties = Breslow;
run;
ods output close;

 /*Efron approximation: option ties = Efron*/
ods output ParameterEstimates = est_ef FitStatistics = fit_ef;
proc phreg data= Tied_events;
       model Time*Censor(0) = Age Ind1 Ind2 Ind3 Ind4 Ind5 Ind6
                             /ties = Efron;
run;
ods output close;

 /*Discrete model: option ties = disctete*/
ods output ParameterEstimates = est_dis FitStatistics = fit_dis;
proc phreg data= Tied_events;
       model Time*Censor(0) = Age Ind1 Ind2 Ind3 Ind4 Ind5 Ind6
                            /ties = discrete;
run;
ods output close;

 /*Random value method: modified time variable*/
ods output ParameterEstimates = est_ran FitStatistics = fit_ran;
proc phreg data= Tied_events;
       model Time_mod*Censor(0) = Age Ind1 Ind2 Ind3 Ind4 Ind5 Ind6
                            /ties = Breslow;
run;
ods output close;
```