STATISTICS IN TRANSITION new series, December 2024
Vol. 25, No. 4, pp. 179–189, https://doi.org/10.59139/stattrans-2024-009
Received – 22.01.2024; accepted – 16.08.2024

Estimation of the Cox model with grouped lifetimes

Piotr Bolesław Nowak[1]

Abstract

This paper presents how random numbers can be used to transform grouped lifetimes into a pseudo-complete sample. The aim of the study is to investigate the Fisher consistency of the partial likelihood estimator of the regression parameters in the Cox model based on the restored sample. It has been proven that for elliptical-type distributional assumptions about explanatory variables the estimators of the regression parameters in the Cox model based on the pseudo-complete sample are consistent up to a scaling factor. A simulation study illustrates the asymptotic properties of the estimates. In addition, real data case analysis is presented.

Key words: Cox model, grouped data, Fisher consistency, elliptical distribution.

1. Introduction

Let T be a random variable denoting survival time and $X = (X_1, \ldots, X_p)^\top$ be a vector of covariates having cumulative distribution function H. The Cox proportional hazard model is a common technique for analysis of censored survival data which assumes that the hazard function of time t, given the covariate value $X = x$ is of the form

$$\lambda(t|x) = \lambda_0(t)\exp(\beta^\top x),$$

where $\lambda_0(t)$ is the baseline hazard function and $\beta \in \mathbb{R}^p$ denotes unknown regression parameters. It implies that the conditional survival function of T given $X = x$ takes the form $S(t|x) = P(T > t|x) = \exp\left(-\Lambda(t)\exp(\beta^\top x)\right)$, where $\Lambda(t) = \int_0^t \lambda_0(s)ds$ is the baseline cumulative hazard function.

Given a random sample $\{(T_i \wedge C_i, (X_{i1}, \ldots, X_{ip}), \delta_i)\}_{i=1}^n$, where $\delta_i = 1(T_i \leq C_i)$ and the censoring variable C is independent of T given the value of $X = x$, Cox (1972) introduced a method of estimating β without considering Λ, which is known as the partial likelihood method. The partial likelihood estimator for the Cox model solves the equation

$$\int \left[y - \frac{\int 1(t \wedge c \geq w) x \exp(\beta^\top x) dF_n(t,c,x)}{\int 1(t \wedge c \geq w)\exp(\beta^\top x) dF_n(t,c,x)} \right] 1(w \leq a)\, dF_n(w,a,y) = 0, \tag{1}$$

where $F_n(t,c,x)$ denotes the empirical distribution function of the random sample and 1 denotes the indicator function.

[1]Institute of Economic Sciences, University of Wrocław, Wrocław, Poland. E-mail: piotr.nowak2@uwr.edu.pl. ORCID: https://orcid.org/0000-0002-7404-2946.

Assume now that time is partitioned into k intervals $A_j = [a_{j-1}, a_j)$, $j = 1, \ldots, k$ and $a_0 = 0$, $a_k = \infty$. For each individual the exact value of X is known but the underlying variable T is unobserved due to grouping mechanism. We only know in what interval each individual died or was censored.

In practice, it is impossible to measure time with infinite precision. For instance, in constructing life tables age is rounded to the nearest year. Moreover, sample elements are often classified into disjoint subsets, like intervals, rectangles, etc. It means that it is not possible to give individual sample values but only the numbers of observations in each specified class (for more examples see Haitovsky (1982)). Inference methods for the grouped survival data can be found in Kalbfleich and Prentice (1973), Thompson (1977), Prentice and Gloeckler (1978). Kalbfleich and Prentice (1973) obtained a generalized linear model with a complementary log-log link function while Thompson (1977) used the logistic model. A comprehensive lecture on discrete hazard models can be found in Fahrmeir and Tutz (2001), in particular, see Chapter 9 for the methods for modelling of discrete survival data. McKeague and Zhang (1996) obtained a Sheppard correction for grouping in the Cox model.

The aim of this paper is to present a different approach from the one mentioned above to estimate the conditional survival function, which we describe in the next section. In the sequel we show that estimating equation (1) can be used for inference about β even when lifetimes coming from the Cox model are grouped into intervals. In the final section, we present simulation study concerning scale Fisher consistency of the proposed estimators and give examples with a real data set.

2. The estimator of the parameters for the grouped Cox model

Since the presented considerations hold also in the case of censoring, in order to simplify notations, we first consider the case without censoring.

Recall, that for grouped data instead of the sample $\{(T_i, (X_{i1}, \ldots, X_{ip})\}_{i=1}^n$ we observe $\{(z_i, (X_{i1}, \ldots, X_{ip})\}_{i=1}^n$, where z_i is a $1 \times k$ vector indicating the grouping interval. Thus, $\sum_{i=1}^n z_i = (n_1, \ldots, n_k)$, where n_i is the total number of deaths in the ith interval.

The estimation of the distribution parameters based on the grouped data is often more difficult than for ungrouped data. For data divided into intervals the most straightforward approach to estimation is to assume that all observations within each finite interval are assigned to its midpoint.

The presented method of estimation in the case of the grouped data is based on the idea that an unobserved lifetime T for given $X = x$ in the interval $A_j = [a_{j-1}, a_j)$ may be replaced by a random variable $\tilde{T}$ generated independently according to some distribution on this set with cumulative distribution function (cdf), namely G_j. Therefore, instead of sample $\{(T_i, (X_{i1}, \ldots, X_{ip})\}_{i=1}^n$ we have $\{(\tilde{T}_i, (X_{i1}, \ldots, X_{ip})\}_{i=1}^n$ and hence the estimating equation (1) can be applied. Throughout this paper we will call this sample the *pseudo-complete* sample generated by the grouped Cox model. The term of the pseudo-complete sample was also used by Whitten et al. (1988) for the restoration of incomplete samples, but their method was applied only to censored samples.

Observe that the density of the random variable $[\tilde{T}|X = x]$ is given by the formula $\tilde{f}(t|x) = \sum_{i=1}^k g_i(t) 1_{A_i}(t) P(T \in A_i|x)$, where g_i is the density function over the set A_i. Now,

denote the conditional survival function of this distribution by $\tilde{S}(t|x)$. From the above description, we conclude that

$$\tilde{S}(t|t \in A_j, x) = P(\tilde{T} > t | t \in A_j, x) = S(a_j|x) + [1 - G_j(t)][S(a_{j-1}|x) - S(a_j|x)].$$

The uniform distribution over $[a_{j-1}, a_j)$ is the most natural choice of G_j for each finite A_j. It corresponds to the piecewise linear approximation of the survival function S. For the last set A_k, it is reasonable to consider shifted exponential distribution or distribution of random variable with probability one at the point a_k. When $1 - G_j(t) = (a + h - t)/h$ is the survival function of the uniform distribution over $[a, a+h)$ and if the interval length h approaches 0, then we have $\tilde{S}(t|t \in [a, a+h), x) \approx S(a|x) + (t-a)S'(a|x)$.

In the next chapter we prove that the described reconstruction of the sample leads to estimators which are consistent up to some positive scale, which is explained below.

3. Scaled Fisher consistency

In statistics, most estimators are defined as solutions to the estimating equations based on the empirical distribution. We say that the estimating equation is Fisher consistent at the model (or in short, the estimator being its solution is Fisher consistent) if the solution to this equation coincides with the true parameter when the empirical distribution is replaced by the true model distribution. For instance, Fisher consistency for the Cox model means that if F_n in (1) is substituted by a joint distribution of (T, C, X), where (T, X) is from the Cox's model with parameter β_0, then $\beta = \beta_0$ is its only solution. Proving Fisher consistency is a primary step in examining the asymptotic properties of M-estimators (see, e.g. Huber and Ronchetti (2009)). This notion was used by Bednarski (1993) in robust method of estimation of regression coefficients based on a modification of partial likelihood estimator.

The scaled Fisher consistency means that solutions to the estimating equation, if the empirical distributions are replaced by the true model distributions, are scaled regression parameters, i.e. $\beta = \alpha\beta_0$ for some scaling factor $\alpha > 0$.

The problem of scaled Fisher consistency for some regression models was considered by Ruud (1983) and Stoker (1986), among others. Another recent important account in such studies is due to Bednarski and Skolimowska-Kulig (2018), who showed that the maximum likelihood estimator for the regression parameters in the classical exponential regression model is scaled Fisher consistent for the extended model. Recently, Bednarski and Nowak (2021), Bednarski, Nowak and Skolimowska-Kulig (2022) have showed that in the Cox model with arbitrary frailty the partial likelihood estimator is also Fisher consistent up to a scaling factor under elliptic type distributional assumptions on explanatory variables.

For further considerations replace the empirical distribution function F_n in (1) by the joint distribution of $(\tilde{T}, X)$, i.e. $\tilde{F}_{\beta_0}(t,x) = \tilde{F}_{\beta_0}(t|x)H(x)$. We always use the subscript β_0 to emphasize that the distribution of (T, X) is under the true value of the parameter β. Thus,

equation (1) becomes

$$\sum_{j=1}^{k}\int_{A_j}\left[y-\frac{\int \tilde{S}_{\beta_0}(w|w\in A_j,x)xe^{\beta^\top x}dH(x)}{\int \tilde{S}_{\beta_0}(w|w\in A_j,x)e^{\beta^\top x}dH(x)}\right]d\tilde{F}_{\beta_0}(w,y)=0. \tag{2}$$

We have the following definition.

Definition 1. *The scaled Fisher consistency of the partial likelihood estimator of β in the Cox model based on a pseudo-complete sample means that equation (2) is satisfied for $\beta=\alpha\beta_0$, where $\alpha>0$ is some scaling factor.*

Remark 1. *Reduction of equation (2)*

Observe that equation (2) can be reduced with an assumption that $EX=0$.

Denoting by μ_0 the expectation of X and after performing some simple algebra this equation can be transformed as follows: $H(x)$ is replaced by $H(x+\mu_0)$ and $\Lambda(w)$ by $e^{\beta_0^\top\mu_0}\Lambda(w)$.

In the view of the above remark our aim is to show that $\tilde{L}(\beta,\beta_0)=0$ is satisfied for $\beta=\alpha\beta_0$, $\alpha>0$, where

$$\tilde{L}(\beta,\beta_0)=\sum_{j=1}^{k}\int_{A_j}\left[\frac{\int \tilde{S}_{\beta_0}(w|w\in A_j,x)xe^{\beta^\top x}dH(x)}{\int \tilde{S}_{\beta_0}(w|w\in A_j,x)e^{\beta^\top x}dH(x)}\right]d\tilde{F}_{\beta_0}(w). \tag{3}$$

The main idea of proving scaled Fisher consistency is based on the construction of an auxiliary function $f_\beta:[0,\infty)\to\mathbb{R}$ defined as follows:

$$f_\beta(\alpha)=\sum_{j=1}^{k}\int_{A_j}\left[\frac{\int(\beta^\top x)\tilde{S}_\beta(w|w\in A_j,x)e^{\alpha\beta^\top x}dH(x)}{\int \tilde{S}_\beta(w|w\in A_j,x)e^{\alpha\beta^\top x}dH(x)}\right]d\tilde{F}_\beta(w). \tag{4}$$

The behavior of the function f_β is described in the following lemma. Its proof is omitted as it is similar to the proof of Lemma 3.1 in Bednarski and Nowak (2021).

Lemma 1. *For any β and any continuous $G_1,\ldots,G_k$ the function f_β has the following properties:*

1. *It is continuous and strictly increasing on $[0,\infty)$.*
2. *$f_\beta(0)<0$.*
3. *$\lim_{\alpha\to\infty} f_\beta(\alpha)>0$.*

Now, let us recall that a p-dimensional random vector X is spherically symmetric distributed if for every orthogonal matrix Γ of size p (i.e. $\Gamma\Gamma^\top=\Gamma^\top\Gamma=I$) the random vector ΓX is distributed as X. Then, the random vector $Y=\mu+AX$ is said to be elliptically symmetric distributed with parameters $\mu\in\mathbb{R}^p$ and covariance matrix Σ_Y, where $\Sigma_Y=AA^\top$. It is known that conditional expectation of Y given $\beta^\top Y=c$ is a linear function with respect to c. In fact, the following lemma can be proved (see also Bednarski and Nowak (2021)).

Lemma 2. *Let Y be a p-dimensional random vector which has an elliptically symmetric distribution with parameters $\mu \in \mathbb{R}^p$ and Σ_Y. Then, for any $\beta \in \mathbb{R}^p$ and any $c \in \mathbb{R}$ it holds*

$$E[Y|\beta^\top Y = c] = \mu + (c - \beta^\top \mu)\frac{\Sigma_Y \beta}{\beta^\top \Sigma_Y \beta}.$$

Now, we are ready to formulate the main theorem, which gives sufficient conditions for the scaled Fisher consistency when the partial likelihood estimator is used for the grouped Cox model based on the pseudo-complete sample.

Theorem 1. *Let the vector of explanatory variables $X = (X_1, \ldots, X_p)^\top$ be elliptically symmetric distributed. Then for any continuous distributions $G_1, \ldots, G_k$ the partial likelihood estimator for the grouped Cox based on the pseudo-complete sample is Fisher consistent up to a scale factor.*

Proof. We show that the equation $\tilde{L}(\alpha\beta_0, \beta_0) = 0$ is satisfied for some scaling factor $\alpha > 0$. Observe that an immediate conclusion from Lemma 1 is that there exists $\alpha_0 > 0$ such that $f_{\beta_0}(\alpha_0) = 0$. Putting $\beta = \alpha_0\beta_0$ we can write the inner integral from the numerator in (3) as the expectation $E(\tilde{S}_{\beta_0}(w|w \in A_j, X)Xe^{\alpha_0\beta_0^\top X})$. Conditioning it on $\beta_0^\top X$ and applying Lemma 2 for X with $\mu = 0$ we have

$$E(\tilde{S}_{\beta_0}(w|w \in A_j, X)Xe^{\alpha_0\beta_0^\top X}) =$$
$$E(E(\tilde{S}_{\beta_0}(w|w \in A_j, X)Xe^{\alpha_0\beta_0^\top X}|\beta_0^\top X)) = \frac{\Sigma_X\beta_0}{\beta_0^\top\Sigma_X\beta_0}E\left((\beta_0^\top X)\tilde{S}_{\beta_0}(w|w \in A_j, X)e^{\alpha_0\beta_0^\top X}\right).$$

Hence, $\tilde{L}(\alpha_0\beta_0, \beta_0) = \frac{\Sigma_X\beta_0}{\beta_0^\top\Sigma_X\beta_0}f_{\beta_0}(\alpha_0) = 0$, which ends the proof. □

Remark 2. *The presence of a censoring variable.*

In the case of the presence of a censoring variable we observe $(T_1 \wedge C_1, X_1, \delta_1), \ldots, (T_n \wedge C_n, X_n, \delta_n)$, where X denotes covariate vector and $\delta = 1(T \leq C)$. Let $F(t,c,x)$ denote the joint distribution of time T, censoring variable C and covariates X under the Cox model. Under the conditional independence of T and C given X one can factorize $dF_{\beta_0}(t,c,x) = dF_{\beta_0}(t|x)dC(c|x)dH(x)$. Now, we replace the random variable $T \wedge C$ by $\tilde{T}$ as follows: when $T \wedge C$ takes the values from A_j then $\tilde{T}$ follows the distribution on the set A_j with cdf G_j on this set. Thus, the pseudo-sample generated by the grouped Cox model consists of $(\tilde{T}_1, X_1, \delta_1), \ldots, (\tilde{T}_n, X_n, \delta_n)$. Then, the Fisher scaled consistency for the grouped Cox model based on the pseudo-complete sample means that the equation $L(\beta, \beta_0) = 0$ is satisfied for $\beta = \alpha\beta_0$, $\alpha > 0$, where

$$L(\beta, \beta_0) =$$
$$\sum_{j=1}^{k}\int_{A_j}\left[y - \frac{\int \tilde{S}_{\beta_0}(w|w \in A_j, x)[1 - C(w|x)]xe^{\beta^\top x}dH(x)}{\int \tilde{S}_{\beta_0}(w|w \in A_j, x)[1 - C(w|x)]e^{\beta^\top x}dH(x)}\right][1 - C(w|y)]d\tilde{F}_{\beta_0}(w|y)dH(y) = 0.$$

From the above it follows that Lemma 1 and Theorem 1 remain applicable in the presence of a censoring variable.

4. Numerical examples

This section presents computational examples for selected distributions and an application of presented method for real and simulation data.

Example 1. (Monte Carlo simulation)

A Monte Carlo experiment for 5000 runs was conducted to investigate properties of the partial likelihood estimation under the pseudo-complete sample when data were generated from the Cox model. The S-Plus programing language was used to generate lifetimes coming from the Cox model. For the true parameter value $\beta_0 = (1, -0.5, 0.5)^\top$, two types of cumulated baseline intensities, $\Lambda(t) = t^{1/2}$ and $\Lambda(t) = t^2$ were used. The vector X was either elliptically distributed with standard normal distributions or non-elliptically distributed with exponential marginals. The sample size was taken $n = 500$ and the grouping was performed for $k = 2, 5, 15, 20$. For each grouping the class $A_k = [a_k, \infty)$ was chosen so that $P_{\beta_0}(T > a_k) = 0.1$ and the group limits $0, a_1, \ldots, a_{k-1}$ were equidistant. After grouping of lifetimes the pseudo-complete samples were created. The uniform distribution on each finite interval and the shifted exponential distribution on the tail were applied. Table 1 shows the results of this experiment.

Table 1: Results of simulation experiment for true parameter $\beta_0 = (1, -0.5, 0.5)^\top$. The first vector in each cell refers to the means of ratios of components of estimates and the true parameters. The second one refers to the standard deviations of the vector estimates of true parameter values.

grouping	Regressors normally distributed		Regressors non-elliptically distributed	
	$\Lambda(t) = t^{1/2}$	$\Lambda(t) = t^2$	$\Lambda(t) = t^{1/2}$	$\Lambda(t) = t^2$
$k = 2$	(3.220, 3.231, 3.214) (0.066, 0.065, 0.065)	(3.105, 3.113, 3.116) (0.065, 0.064, 0.066)	(6.262, 2.382, 4.255) (0.045, 0.041, 0.042)	(6.288, 2.386, 4.257) (0.045, 0.041, 0.044)
$k = 5$	(2.056, 2.060, 2.053) (0.068, 0.066. 0.066)	(1.216, 1.213, 1.210) (0.075, 0.070, 0.069)	(3.500, 1.848, 2.492) (0.054, 0.041, 0.047)	(1.557, 1.420, 1.343) (0.090, 0.045, 0.056)
$k = 15$	(1.544, 1.542, 1.545) (0.073, 0.066, 0.066)	(1.089, 1.091, 1.088) (0.079, 0.072, 0.072)	(2.465, 1.624, 1.829) (0.067, 0.042, 0.051)	(1.098, 1.223, 1.073) (0.075, 0.047, 0.051)
$k = 20$	(1.462, 1.463, 1.468) (0.074, 0.069, 0.068)	(1.084, 1.087, 1.081) (0.077, 0.072, 0.071)	(2.292, 1.581, 1.732) (0.071, 0.042, 0.053)	(1.066, 1.203, 1.054) (0.068, 0.048, 0.050)

Simulations indicate good asymptotic performance of the estimator under normally distributed covariates. Note, that each elliptically distributed vector X can be chosen as a member of a large family of probability distributions like multivariate normal distributions, multivariate t-distributions, multivariate Logistic and Laplace distributions and many others. For this case components of the first vectors in each cell are almost the same, which

shows that we have the estimation of the regression parameter up to the same scaling factor. It is interesting that even for grouping for $k = 2$ we can estimate the regression parameter up to a scaling constant which is approximately equal to 3.2 and 3.1 for $\Lambda(t) = t^{1/2}$ and $\Lambda(t) = t^2$, respectively. The scaling factors decrease as the number of classes increase. For grouping with $k = 15, 20$ and $\Lambda(t) = t^2$ a scaling factor is near one, which corresponds to the consistent estimation of the regression parameter. On the other hand, we observe bad performance of the estimators under departure from the elliptical type distributional assumption of explanatory variables when the number of grouping classes is small, especially for $k = 2$. As the number of classes increase the estimators may approach to the true parameter despite non-elliptically distributed regressors, for instance, see the case for $k = 15$ and $\Lambda(t) = t^2$, where the estimation seems to be correct.

Example 2. (Life table)

Another example presented here compares two estimation methods for the life table for gender and race (see Table 2 based on article by Arias (2007)). These data were also considered by Agresti (2010) on page 127.

The first method of the estimation is applied in order to reconstruct the entire sample from the grouped sample using random numbers generated according the uniform distribution on each interval. We assumed that $A_7 = (95, 120)$, because according to the International Database on Longevity the longest-lived person ever form the United States died at the age of 119 years and 97 days, see also Kestenbaum and Ferguson (2010).

For two explanatory variables, gender g (1 = female; 0 = male) and race r (1 = black; 0 = white), the Cox model was fitted to the pseudo-complete sample of size 1000 for each of the four groups.

As a second model, we used the generalized linear model (GLM) with complementary log-log link function, i.e.

$$\log(-\log(1 - P(Y \leq j))) = \theta_j + \beta_1 g + \beta_2 r, \; j = 1, 2, \ldots, 6.$$

Table 2 contains fitted distributions, the first value in each parenthesis corresponds to the Cox model and the second value in each parenthesis to the GLM. For each of the four distributions and for each of the estimation methods, we calculated the dissimilarity index, which is the half the sum of absolute differences between the fitted and estimated population distributions. This index takes values (in percent) 2.2, 6.6, 7.2, 3.5 for the Cox model and 2.7, 6.8, 6.8, 3.3 for the GLM. The differences in estimates of two mentioned methods are very small.

Example 3. (Veteran data)

The next example compares the two estimation methods for the Veteran's Administration lung cancer data, see Kalbfleisch and Prentice (1980). This data set is frequently used to test different estimation. There were continuous covariates: Karnofsky rating, disease duration and age whereas binary ones are prior therapy (yes=1 or no=0), treatment (standard=1 or test=0) and four cell types (squamous, small, large and adeno). Because of colinearity of these cell types, we take into consideration in this model only three of them, namely squamous, small and adeno.

Table 2: Observed and fitted (in parentheses) life-length distributions of U.S. residents, as percentages. The first value in each parenthesis corresponds to the Cox model based on pseudo-complete sample, the second one to the GLM with complementary log-log link function.

Gender	Race	Life Length						
		0-20	20-40	40-50	50-65	65-80	80-95	over 95
Female	Black	1.8 (1.5, 1.5)	2.4 (2.6, 2.6)	3.7 (3.4, 3.3)	12.9 (12.6, 12.4)	29.6 (30.0, 29.9)	39.3 (40.8, 41.5)	10.3 (9.1, 8.8)
	White	0.9 (1.2, 1.2)	1.3 (2.2, 2.0)	1.9 (2.7, 2.6)	8.0 (10.4, 9.9)	25.9 (26.3, 25.3)	49.7 (43.1, 43.5)	12.3 (14.1, 15.5)
Male	Black	2.6 (2.1, 2.2)	4.9 (3.6, 3.8)	5.6 (4.6, 4.8)	20.2 (16.6, 17.3)	34.7 (35.2, 36.1)	27.8 (34.5, 33.2)	4.2 (3.4, 2.6)
	White	1.3 (1.7, 1.7)	2.8 (2.9, 2.9)	3.2 (3.7, 3.8)	12.2 (13.9, 14.0)	32.8 (32.1, 32.2)	42 (39.2, 39.3)	5.7 (6.5, 6.1)

In order to present the result for the pseudo-complete sample we grouped lifetimes into twenty equidistance classes, i.e. $k = 20$. The range of lifetime is 1–999. Each grouped lifetime was replaced by a random number according to the uniform distribution on the corresponding interval. Scaled values of estimates for complete and pseudo-complete sample are presented in Table 3. The differences in scaled estimates are very small, i.e. the maximum absolute difference is no more than 0.05.

Table 3: Comparison of partial likelihood estimation for complete and pseudo-complete sample for the Veteran's Administration lung cancer data.

	complete sample			pseudo-complete sample		
covariates	ple	scaled ple	p-value	ple	scaled ple	p-value
karnofsky	-0.0328	-0.0314	0.0000	-0.0327	-0.0299	0.0000
diag time	0.0001	0.0001	0.9929	-0.0040	-0.0036	0.6683
age	-0.0087	-0.0083	0.3492	-0.0015	-0.0014	0.7807
prior	0.0072	0.0068	0.7579	0.0018	0.0017	0.8328
squamous	-0.4013	-0.3839	0.1557	-0.4072	-0.3720	0.1574
small	0.4603	0.4403	0.0838	0.4404	0.4024	0.1105
adeno	0.7948	0.7604	0.0087	0.8498	0.7764	0.0076
tratment	0.2946	0.2818	0.1558	0.3390	0.3098	0.1131

Example 4. (Estimation in the Cox model with rounded data)

Let us recall that the Cox model is based on several restrictive assumptions and one of them says that there were no tied values among the observed survival times. When constructing a new partial-likelihood function we must assume that the roundings for particular survival times appear by imprecision in the measurements of survival time. Therefore, when we have d values rounded to the one value, in fact they could have been observed in any of the $d!$ possible orders. The exact form of the partial-likelihood function is obtained by modification of the partial-likelihood function to include all possible arrangements. Then, we get expressions inconvenient for further calculations, therefore we use approximations.

Approximation, both introduced by Breslow (1974) and Efron (1977), provides simpler expressions than an exact function, but still include effect of rounded data.

In order to apply randomization procedure to rounded data we replace each tied survival time, namely t, by a random variable according to uniform distribution on the interval $(t-\varepsilon, t+\varepsilon)$, $\varepsilon > 0$.

Now, we illustrate this randomization procedure by considering data on HMO examination of patients infected with HIV (see Hosmer and Lemeshow (1999)), where 100 patients participated in the study, with 31 different survival times. The number of people with the same time survival rates ranged between 1 and 17. For the simulation, we assumed that $\varepsilon = 0.5$.

Table 4: Comparison of estimation results for different estimation methods.

Method	AGE		DRUG	
	Coeff.	Sd.Err.	Coeff.	Sd.Err.
Exact	0.0977	0.0187	1.0226	0.2572
Breslow	0.0915	0.0185	0.9414	0.2555
Efron	0.0971	0.0186	1.0167	0.2562
Random	0.0976	0.0186	1.0307	0.2577

Table 4 shows estimation results for methods mentioned above. Note that using the randomization method (see the last row of Table 4), we get the results that are very close to the exact one. In fact, estimators calculated by all four methods are close to each other and their standard errors are almost identical.

5. Final conclusions

The Cox model is based on several restrictive assumptions and one of them assumes continuous survival time. This assumption may not be fulfilled in many situations, e.g. when the data are rounded and then at least two events may occur at one point in time. Another situation concerns the case when survival data are grouped. In general, data grouping is a frequently used data presentation mechanism in practical applications. There are well-known methods of inference based on grouped data in the statistical literature, but due to data compression, the resulting estimators may be less effective or more biased than those obtained on the basis of the full sample.

This paper presents a method of estimation of the regression parameters in the Cox model, when lifetimes are grouped into a set of intervals. We showed how using random numbers, which are easily available in statistical packages, one can obtain a reconstruction of a simple sample, called here a pseudo-complete sample, and hence the classic Cox estimator can be still used. We noticed that in case of using a uniform distribution on each grouping interval, the described randomization method corresponds to the approximation of the survival function by a piecewise linear function. We proved that for the pseudo-complete sample the partial likelihood method of estimation leads to the consistent estimation of regression parameters up to a scaling factor if covariates are elliptically distributed. By stan-

dard asymptotic argumentation it means that solutions to equation (1) for restored samples converge to scaled regression parameters as the sample increases and they are asymptotically normal.

The problem discussed in this paper is important because initial data are often aggregated and then classical methods based on the assumption of continuity of the dependent variable are limited. Therefore, the presented randomization method can also be used in other regression models, where a dependent variable is grouped.

Acknowledgements

The author thanks the referees for comments leading to important improvements in the paper.

References

Agresti, A., (2010). *Analysis of Ordinal Categorical Data, 2nd Edition*, John Wiley and Sons.

Arias, E., (2007). United States life tables, 2004. *National Vital Statistics Reports*, 56(9), pp. 1–40.

Bednarski, T., (1993). Robust estimation in Cox's regression model. *Scandinavian Journal of Statistics*, 20(3), pp. 213–225.

Bednarski, T., Nowak, P. B., (2021). Scaled Fisher consistency of partial likelihood estimator in the Cox model with arbitrary frailty. *Probability and Mathematical Statistics*, 41(1), pp. 77–87.

Bednarski, T., Nowak, P. B., Skolimowska-Kulig, M., (2022). Scaled Fisher consistency for the partial likelihood estimation in various extensions of the Cox model. *Statistics in Transition new series*, 23(2), pp. 185–196.

Bednarski, T., Skolimowska-Kulig, M., (2018). Scaled consistent estimation of regression parameters in frailty models. *Acta Universitatis Lodziensis. Folia Oeconomica*, 5(338), 133–142.

Breslow, N., (1974). Covariance analysis of censored survival data. *Biometrics*, 30(1), pp. 89–99.

Cox, D. R., (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B*, 34(2), pp. 187–220.

Efron, B., (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 71(359), pp. 557–565.

Fahrmeir, L., Tutz, G., (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd Edition*, Springer-Verlag, New York.

Haitovsky, Y., 1982. Grouped data, in *Encyclopedia of Statistical Sciences 3*, John Wiley and Sons, pp. 527–536.

Hosmer, D., Jr, Lemeshow, S., (1999). *Applied Survival Analysis. Regression Modeling of Time to Event Data*, John Wiley and Sons.

Huber, P. J., Ronchetti, E. M., (2009). *Robust Statistics, 2nd Edition*, John Wiley and Sons.

Kalbfleich, J. D., Prentice, R. L., (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 60(2), pp. 267–278.

Kalbfleich, J. D., Prentice, R. L., (1980). *The Statistical Analysis of Failure Time Data*, John Wiley and Sons.

Kestenbaum, B., Ferguson, R., (2010). Supercentenarians in the United States, in *H. Maier et al. (eds.), Demographic Research Monographs*, Springer, Berlin, Heidelberg, pp. 43–58.

McKeague, I. W., Zhang, M. J., (1996). Fitting Cox's Proportional Hazards Model Using Grouped Survival Data, in *N.P. Jewell et al. (eds.), Lifetime Data: Models in Reliability and Survival Analysis*, Kluwer, Boston, pp. 227–232.

Prentice, R. L., Gloeckler, L. A., (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, 34(1), pp. 57–67.

Ruud, P., (1983). Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models. *Econometrica*, 51(1), pp. 225–228.

Stoker, T., (1986). Consistent estimation of scaled coefficients. *Econometrica*, 54(6), pp. 1461–1481.

Thompson, W. A., (1977). On the treatment of grouped observations in live studies. *Biometrics*, 33(3), pp. 463–470.

Whitten, B. J., Cohen, A. C., Sundaraiyer, V., (1988). A pseudo-complete sample technique for estimation from censored samples. *Communications in Statistics - Theory and Methods*, 17(7), pp. 2239–2258.