EKONOMETRIA
ECONOMETRICS

3(45)•2014

Publishing House of Wrocław University of Economics
Wrocław 2014

Copy-editing: Marcin Orszulak

Layout: Barbara Łopusiewicz

Proof-reading: Barbara Cibis

Typesetting: Małgorzata Czupryńska

Cover design: Beata Dębska

This publication is available at www.ibuk.pl, www.ebscohost.com,
Lower Silesian Digital Library www.dbc.wroc.pl,
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
and in The Central and Eastern European Online Library www.ceeol.com,
as well as in the annotated bibliography of economic issues of BazEkon http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Information on submitting and reviewing papers is available on
the Publishing House's website
www.wydawnictwo.ue.wroc.pl

ISSN 1507-3866

The original version: printed
Printing: EXPOL, P. Rybiński, J. Dąbek, sp.j.
ul. Brzeska 4, 87-800 Włocławek

Contents

Streszczenia

EKONOMETRIA ECONOMETRICS 3(45) • 2014

ISSN 1507-3866

Jadwiga Borucka
Warsaw School of Economics
e-mail: jadwiga.borucka@gmail.com

EXTENSIONS OF COX MODEL FOR NON-PROPORTIONAL HAZARDS PURPOSE

Summary: Cox proportional hazard model is one of the most common methods used in time to event analysis. The idea of the model is to define a hazard level as a dependent variable which is explained by the time-related component (so-called baseline hazard) and the covariates-related component. The model is based on several restrictive assumptions one of which is the assumption of proportional hazard. However, if this assumption is violated, this does not necessarily prevent an analyst from using Cox model. The current paper presents two ways of model modification in the case of non-proportional hazards: introducing interactions of selected covariates with function of time and stratification model. Calculations performed give the evidence that both methods result in better model fit as compared with the original model. Additionally, they allow interpreting the parameters estimates more precisely, taking into account the effect of the covariate at the hazard level that is changing over time. The choice of the appropriate method of tied events handling however is not straightforward and should be adjusted to the particular analysis purpose.

Keywords: Cox model, survival analysis, non-proportional hazards.

DOI: 10.15611/ekt.2014.3.07

1. Introduction

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

$$\lambda(t, \boldsymbol{x}) = \lambda_0(t) \exp(\beta \boldsymbol{x}), \tag{1}$$

where: (t, x) – hazard function that depends on timepoint t and vector of covariates $\boldsymbol{x}$,
$\lambda_0(t)$ – baseline hazard function that depends on time only,
$\exp(\beta \boldsymbol{x})$ – covariates-related component.

Cox model is based on several restrictive assumptions. One of them is the assumption of proportional hazard that the name of the model refers to and which results directly from the model formula as follows:

$$HR = \frac{\lambda(t, \boldsymbol{x}_1)}{\lambda(t, \boldsymbol{x}_2)} = \frac{\lambda_0(t)\exp(\beta \boldsymbol{x}_1)}{\lambda_0(t)\exp(\beta \boldsymbol{x}_2)} = \frac{\exp(\beta \boldsymbol{x}_1)}{\exp(\beta \boldsymbol{x}_2)} = \exp[\beta(\boldsymbol{x}_1 - \boldsymbol{x}_2)], \quad (2)$$

where: *HR* – hazard ratio,

$\boldsymbol{x}_1$ – vector of covariates of subject I,

$\boldsymbol{x}_2$ – vector of covariates of subject II.

The assumption states that the hazard ratio for two subjects who are characterized by different sets of covariates depends only on the values of these covariates and does not depend on time. In other words: the hazard ratio is constant over time which means that the effect of a given covariate on a hazard level is the same at all timepoints. There are various opinions on the importance of this assumption with regard to parameters interpretation. Some authors state that its violation is nothing extremely problematic as in such cases the parameter for a covariate for which the assumption is not satisfied can be understood as "average effect" over timepoints that are observed in a dataset [Allison 1995]. The others, however, underline the importance of this assumption [Hosmer, Lemeshow 1999] and suggest potential modification of the model if the hazard ratio turns out not to be constant over time for some covariates. While in some situations measuring "average effect" of a covariate for which the proportional hazard assumption is not satisfied might be enough, it is possible to recall cases where this approach is not sufficient. Hosmer and Lemeshow [1999] discuss this issue and give an example of randomized clinical trials in which a study site is relatively often used as a covariate in Cox model. Such an approach results in assuming that baseline hazards are proportional across study sites which might not necessarily be justified. In such cases it would be worth taking this fact into account and estimate the model adjusting for potentially time-varying effect of a study site rather than stating that parameter estimate for a site expresses its "average effect" at the hazard level.

There are several methods that enable verification of the proportional hazard assumption. Firstly, one can consider a graphical method which is based on the plot of "log-negative-log" of the Kaplan-Meier estimator of the survival function presented separately for each group defined on the basis of the values of a covariate for which the assumption is verified [Hosmer, Lemeshow 1999]. If the assumption is satisfied, the plot should present several curves with the distance between them that does not change over time. One possible disadvantage of this method might be the fact that visual assessment of how far these lines are from parallel position might be a problem, especially for small samples.

The other graphical method employs Schoenfeld residuals which are expected to have average equal to 0, which might be assessed on the basis of the plot (it is

expected that residuals will show no trend over time). The third method adds to the model interaction of the covariate of interest with time – if such a variable turns out to be statistically significant, it indicates that the proportional hazard assumption might be violated. The last method is not only a way to verify the assumption, but also a potential solution to the problem of its violation which will be discussed more thoroughly in the next section.

As soon as it is stated that the proportional hazard assumption is not satisfied for a covariate, one should decide which approach is to chose. As mentioned before, one can think of the parameter estimate as of the average strength of the covariate impact on the hazard rate. If this is the case, nothing more should be done in terms of Cox model construction. If the influence, however, varies over time and this changing impact is also of interest, then one of the available methods of Cox model modification for non-proportional hazards might be applied. There are two methods that are considered the most often: adding a covariate to the model which is defined as an interaction of a particular covariate with a function of a time variable and the stratification model. The next two sections present these methods in more detail, including practical example application on the dataset containing data for 60 subjects from an open-label clinical trial: age at screening (in years), site (coded as SITE = 1 that corresponds to site B or SITE = 2 that corresponds to site A) and time from the beginning of study to death/censoring (in days) as well as censoring information (coded as CENSOR = 1 for subjects who experience the event, here: death, and CENSOR = 0 otherwise). Cox model is used in order to analyze time to death among subjects enrolled in the study with regard to the age of patients and the study site. For presenting purposes a simple Cox model is used including age and site as explanatory variables. The calculations are performed in SAS® Base 9.3. Graphs are plotted in Microsoft Excel on the basis of the results obtained in SAS®; however, it is also possible to get graphs directly from SAS®. All relevant SAS® codes are included in Appendix.

2. Interaction with time

The first method uses interactions with time for covariates for which the assumption is not satisfied. This method is in fact both the way to identify such covariates in the model and solution of the problem at the same time [Allison 1995]. After adding an interaction of some function of a time variable with a covariate included in the initial model, the statistical significance is verified. If a newly added variable turns out to be significant, it indicates that the proportional hazard assumption is not satisfied for a given covariate, which means that its effect is changing over time. Including the interaction in the model enables interpretation of the parameters taking into account the fact that the covariate's influence at the hazard level is not constant. As far as the function type is concerned, some authors suggest using a logarithm rather than any other function [Quantin et al., 1996], the others, however, underline that

there is no theoretical reason to choose a logarithm as this approach is seen rather as a technical solution which allows avoiding numerical problems [Allison 1995]. PHREG procedure is robust to described problems [Allison 1995]; thus, a simple linear function is chosen.[1]

The sample dataset is used to estimate Cox proportional model, where the event is defined as death of a patient, time is measured in days from the beginning of the open-label study, age and site are included as covariates. The exact method according to Kalbfleisch and Prentice [1980] is used in the model estimation in order to account for presence of tied events in the dataset. SAS® Code 1 included in Appendix estimates the model, using TIME as a time variable, CENSOR as a censoring indicator (0 means lack of event), AGE and SITE as covariates.

Table 1. Model 1: Cox model including AGE and SITE as covariates

Variable	Parameter Estimate	SE	Chi-square	*p*-value	Hazard Ratio
AGE	0.20690	0.07405	7.8069	0.0052	1.230
SITE	−0.74290	0.38926	3.6423	0.0563	0.476

Source: own calculations by means of SAS® Base 9.3.

The convergence criterion is satisfied, both covariates are significant at the significance level of 0.1. The proportional hazard assumption is verified for both variables, using interactions with time and Schoenfeld residuals plots. For a categorical variable SITE, an additionally plot of "log-negative-log" is presented.

Proportional hazard assumption verification for SITE:

(a) Plot of "log-negative-log" of survival function:

The estimates of the survival function are obtained by means of LIFETEST procedure with STRATA statement, as stated in SAS Code 2 in Appendix.

[1] In the analyzed example two approaches were considered: adding interaction of the logarithm of time with the covariate of interest, as suggested by Quantin et al. [1996] as well as using a simple linear function of time, as proposed by Allison. For each of the covariates in the model, AGE and SITE, the model was re-estimated using each of the functions listed above. In terms of the statistical significance of the interaction with time, both types of function led to the same conclusions. The analysis of the hazard ratio over time was performed for SITE covariate only if no statistically significance for interaction with AGE was revealed. The plots of the hazard ratio against time performed on the basis of both approaches (using the logarithm of time and the simple linear function of time in the interaction with SITE) were very similar to each other; thus, the choice between the linear and the logarithm function in this particular example does not have a significant impact on the results. In this case the results obtained with the use of the simpler linear function are presented in the further part of the article.

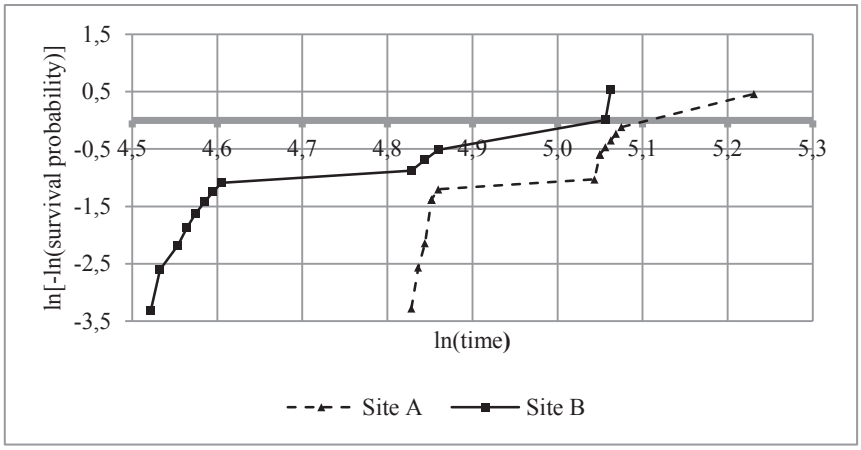

Figure 1. "Log-negative-log" of the survival function

Source: own calculations by means of SAS® Base 9.3

(b) Plot of Schoenfeld residuals:

In order to obtain Schoenfeld residuals plots, it is necessary to estimate Cox model using PROC PHREG and save Schoenfeld residuals in a separate dataset, adding OUTPUT OUT statement. The GPLOT procedure might be used to generate the plot of residuals as a function of time (see SAS Code 3 in Appendix).

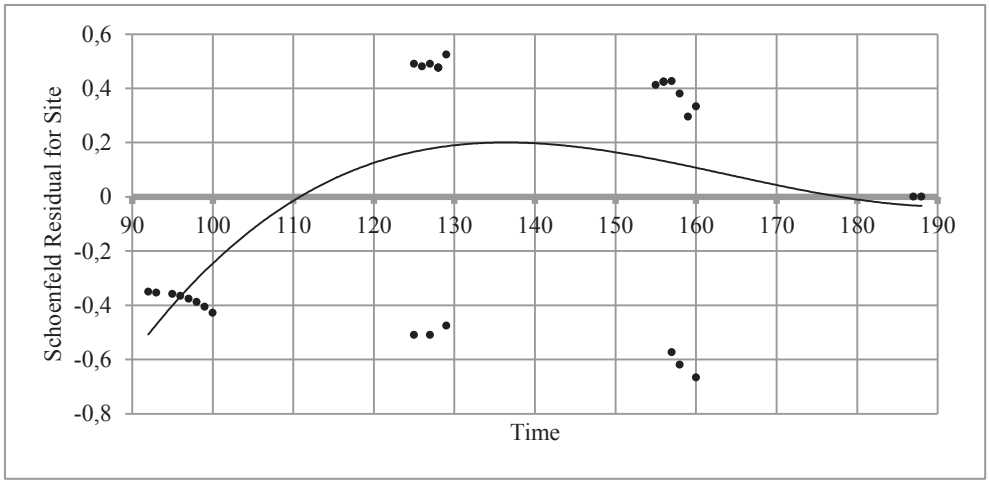

Figure 2. Schoenfeld residuals plot for SITE

Source: own calculations by means of SAS® Base 9.3.

(c) Interaction of SITE with TIME variable:

Adding interaction with time to Cox model is quite simple in PHREG procedure. The additional variable needs to be named in the MODEL statement and then defined using programming statements available in the procedure (see SAS Code 4 in Appendix).

Table 2. Model 2: Cox model including AGE and SITE as covariates and SITE by TIME interaction

Variable	Parameter Estimate	SE	Chi-square	*p*-value	Hazard Ratio
AGE	0.23313	0.07399	9.9268	0.0016	1.263
SITE	−5.90164	2.80362	4.4311	0.0353	0.003
SITE*TIME	0.03985	0.02123	3.5233	0.0605	1.041

Source: own calculations by means of SAS® Base 9.3.

Proportional hazard assumption verification for AGE:

(a) Plot of Schoenfeld residuals:

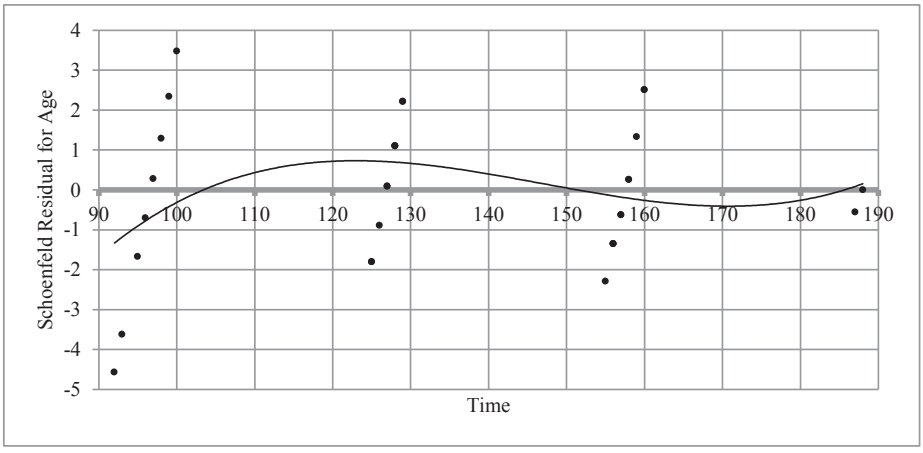

Figure 3. Schoenfeld residuals plot for AGE

Source: own calculations by means of SAS® Base 9.3.

(b) Interaction of AGE with TIME variable:

Table 3. Model 3: Cox model including AGE and SITE as covariates and AGE by TIME interaction

Variable	Parameter Estimate	SE	Chi-square	*p*-value	Hazard Ratio
AGE	0.01692	0.38357	0.0019	0.9648	1.017
SITE	−0.70788	0.39323	3.2406	0.0718	0.493
AGE*TIME	0.00149	0.00296	0.2516	0.6160	1.001

Source: own calculations by means of SAS® Base 9.3.

On the basis of the presented results, it can be stated that the proportional hazard assumption seems to be satisfied for AGE – the interaction of AGE and TIME is not statistically significant at any acceptable level, Schoenfeld residuals on the plot do not show any trend, smoothed line has approximate zero slope. For SITE, however, the conclusion seems to be quite opposite – the interaction with time is statistically significant at the significance level 0.1. The plot of Schoenfeld residuals does not give a straightforward answer; however, it might suggest that hazards are not proportional across study sites, which can be stated also on the basis of "log-negative-log" as distance between the lines is not constant for all values of TIME. This might lead to the conclusion that the proportional hazard assumption is violated for SITE variable and the effect of this variable might be changing over time. Thus, it would be worth considering the model including interaction of SITE and TIME. Comparing Model 1 (initial one) and Model 2 (with TIME by SITE interaction included), it can be seen that information criteria have lower values in Model 2. The likelihood ratio test in this case is in fact a test for the significance of TIME by SITE interaction, as it is the only covariate added, and leads to the rejection of the null hypothesis which assumes that an added variable is not significant. It seems then that adding a new variable to the model resolves the problem with the violated assumption and improves fit statistics. Thus, considering two models presented earlier, Model 2 should be chosen rather than Model 1. Let us focus on the parameters interpretation for SITE variable in both models. As far as Model 1 is concerned, interpretation is quite straightforward. Parameter estimate equals to −0.7429 results in the hazard ratio for the binary variable at the level of 0.48, which means that the subjects from site A are approximately 52% less likely to die than the subjects from site B. Due to the fact that the proportional hazard assumption is violated for SITE, this interpretation might only refer to "average" effect of SITE, as suggested by Allison. Now, let us take into account the fact that the effect of this covariate is changing over time. In order to calculate the hazard ratio between site A and site B on the basis of Model 2 estimation, the following equation is derived:

$$HR = \frac{\lambda(t, age = a, site = 2)}{\lambda(t, age = a, site = 1)} = \\ = \frac{\lambda_0(t)\exp[\beta_1 a + 2\beta_2 + 2\beta_3 \mathrm{t}]}{\lambda_0(t)\exp[\beta_1 a + \beta_2 + \beta_3 \mathrm{t}]} = \exp[\beta_2 + \beta_3 \mathrm{t}], \tag{3}$$

where: β_1 – parameter estimate for age,
β_2 – parameter estimate for site,
β_3 – parameter estimate for interaction of site and time.

As it can be seen, the hazard ratio depends on time; thus, the interpretation of the site effect at the hazard level should incorporate the value of time variable as well. The plot in Figure 4 presents how the hazard ratio between two subjects of the same age, but from different sites changes over time.

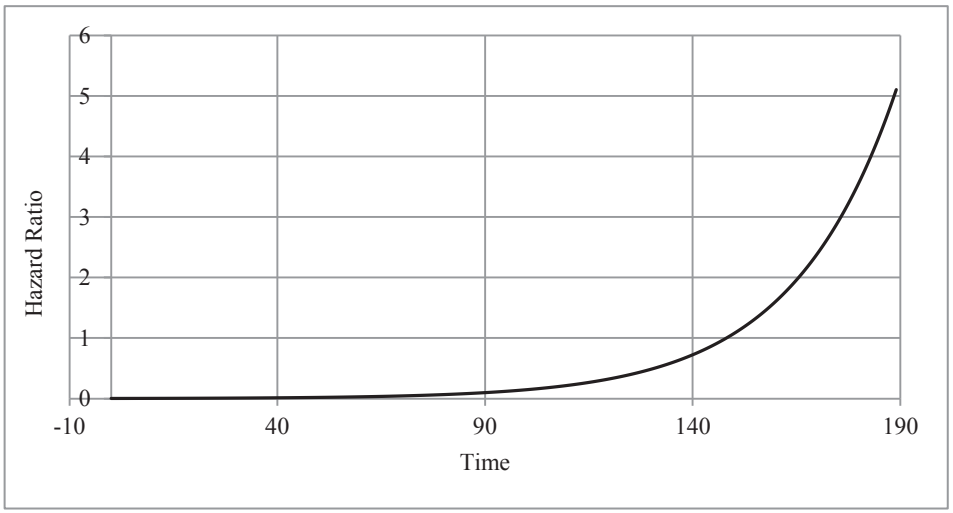

Figure 4. Hazard ratio for SITE over time

Source: own calculations be means SAS® Base 9.3.

On the basis of the plot, it can be stated that the hazard ratio is very low up to the 80th day, which means that for relatively low survival time the subjects from site B are much more likely to die than the subjects from site A; however, the hazard ratio constantly increases so this difference is becoming smaller and smaller. On the 130th day the subjects from site A are approximately 50% less likely to die than the subjects from site B. The hazard ratio reaches the value of 1 on the 148th day, which means that the chances of dying are equal for the subjects treated in both sites. After that time the hazard ratio rapidly increases and exceeds 3 after the 176th day, which means that eventually the subjects from site A are even three times more likely to die than the subjects treated in site B. It should be mentioned that the first event in the whole sample was recorded on the 92nd day; thus, drawing conclusions for the period 0–92 days might not be reliable and the interpretation should be focused rather on the values of TIME greater than or equal to 92. In general, it might be stated that the subjects from site B seem to experience the event relatively early, but if they survive long enough, they are not likely to die in a later phase. The subjects from site A are more likely to live longer, but after some time point they seem to experience the event approximately as often as the subjects from site B. To be more specific, 14 subjects from site B are dying between the 92nd and the 160th day, 17 subjects who die in site A have survival times between the 125th and the 188th day in the study. As compared with the results from Model 1, where it is stated that the subjects from site A are "on average" approximately 52% less likely to experience the event, it is worth mentioning that after accounting for the fact that the effect of SITE is not constant over time, this difference turns out to be much higher at some time points and – what

is more important – the hazard ratio is lower than 1 in the early phase (meaning that the subjects from site B are a more risky group), but exceeds 1 after the 148th day, which indicates that the subjects from site A become more likely to die.

3. Stratified model

The second method that allows handling non-proportional hazards is stratification. The main idea is to split the whole sample into subgroups on the basis of categorical variable which is called a stratification variable and subsequently re-estimate the model letting the baseline hazard function differ between these subgroups. It makes sense to choose a categorical covariate as a stratification variable if it interacts with time (i.e. the proportional hazard assumption is not satisfied for this covariate) and is not of primary interest as the stratification of the model automatically excludes a stratification variable from an explanatory variable set. It should be noted that stratification might be also performed on the basis of a continuous covariate; however, in this case creating a grouped version of an original variable is required. As far as the stratification variable is concerned, it is assumed that this is a covariate that has an impact on the outcome but the estimation of its effects is not crucial. A stratification variable needs to allow splitting a sample into subgroups which are internally homogenous and differ from each other in terms of the baseline hazard function. The effects of a stratification variable, which cannot be directly estimated in a stratified model, are incorporated within the group-specific baseline hazard function. In the literature, it is suggested that stratification variable should be fixed by design or identified on the basis of the previous research [Hosmer, Lemeshow 1999].

As far as coefficients estimates are concerned in the basic form of the stratified model, it is assumed that they are constant across strata groups. However, it is possible to include the interaction of a stratification variable and another covariate in order to take into account different slopes [Hosmer, Lemeshow 1999]. In general, the stratified model for stratum s is defined as follows:

$$\lambda_s(t, \boldsymbol{x}) = \lambda_{s0}(t) \exp(\beta \boldsymbol{x}), \tag{4}$$

where $s = 1, 2, ..., S$ and S is the total number of subgroups created on the basis of a stratification variable. The partial likelihood function formula is similar to the function proposed by Cox having an additional subscript indicating the stratum number. By multiplying the partial likelihood function for each stratum, the full partial likelihood function is obtained (for details refer to Hosmer, Lemeshow [1999]). After the stratified model is estimated, it is possible to obtain the estimation of baseline survival and baseline cumulative hazard functions for each stratum. Additionally, one can estimate covariates-adjusted survival and cumulative hazard functions. There is available option BASELINE in PHREG procedure which allows

obtaining these estimates for each stratum. Calculations are performed for each set of covariates that are specified by the user. If no input dataset containing specified sets of covariates is defined, then SAS calculates the survival and the cumulative hazard function for each value of a stratification variable, taking the average of continuous variables within each stratum.

Let us consider SITE for which it is known that the proportional hazard assumption is violated as a stratification variable. It will be not possible to obtain parameter estimates for SITE; however, using BASELINE statement enables estimating survival and cumulative hazard function for each site separately, adjusting for age. Cox model is re-estimated in the modified formula, using STRATA statement in PHREG procedure (see SAS Code 5 in Appendix).

Table 4. Model 4: Cox model including AGE as a covariate and SITE as a stratification variable

Model 4					
Stratum	Site	Total	Event	Censored	% Censored
1	Site A	30	17	13	43.33
2	Site B	30	14	16	53.55
Variable	Parameter Estimate	SE	Chi-square	*p*-value	Hazard Ratio
AGE	0.20959	0.07534	7.7389	0.0054	1.233

Source: own calculations by means of SAS® Base 9.3.

The results of the model estimation are presented in Table 4. The convergence criterion is satisfied, AGE – which is the only covariate in the stratified model – is significant at any acceptable level. As one can notice, the hazard ratio for AGE does not differ to a large extent as compared with Model 1 and is equal to 1.23, which means that every year the risk of dying increases by 23% as compared with the previous year. However, it should be taken into account that the hazard ratio calculated by SAS concerns the comparison of the subjects with one year age difference. While comparing the subjects with larger age difference, e.g. 20 years, hazard ratios are equal to 66 for the stratified model and to 63 for the initial model, which makes the difference between the initial model and the stratified model more visible. Additionally, the plots of the cumulative hazard and the survival function are presented in Figure 5 and 6 (see SAS Code 6 in Appendix).

As expected on the basis of the previous results, the subjects from site A tend to have relatively longer survival times as compared with the subjects from site B. The cumulative hazard function for almost all time points is higher for site B, which means that the expected number of events till the given time point is usually higher for subjects treated in this site.

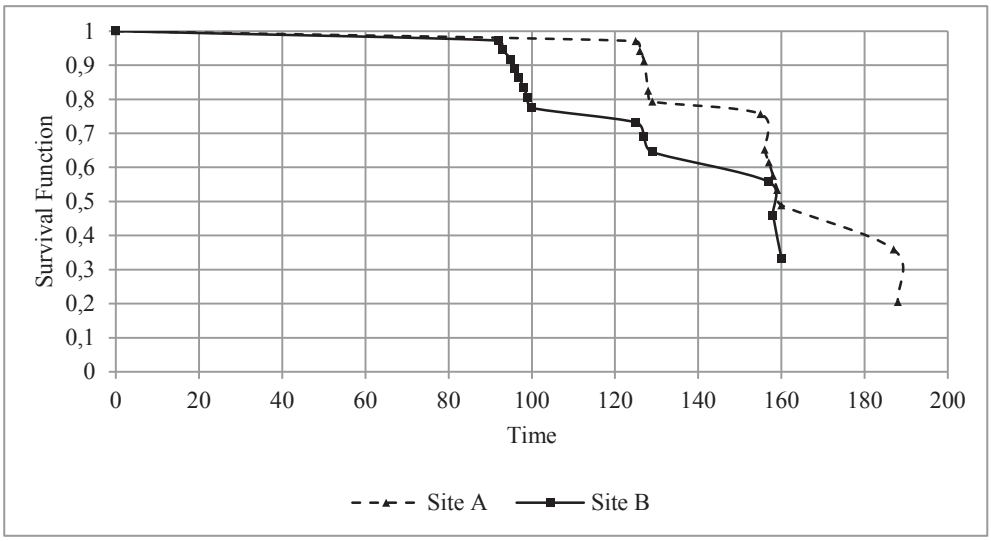

Figure 5. Plot of the survival function by SITE

Source: own calculations by means of SAS® Base 9.3.

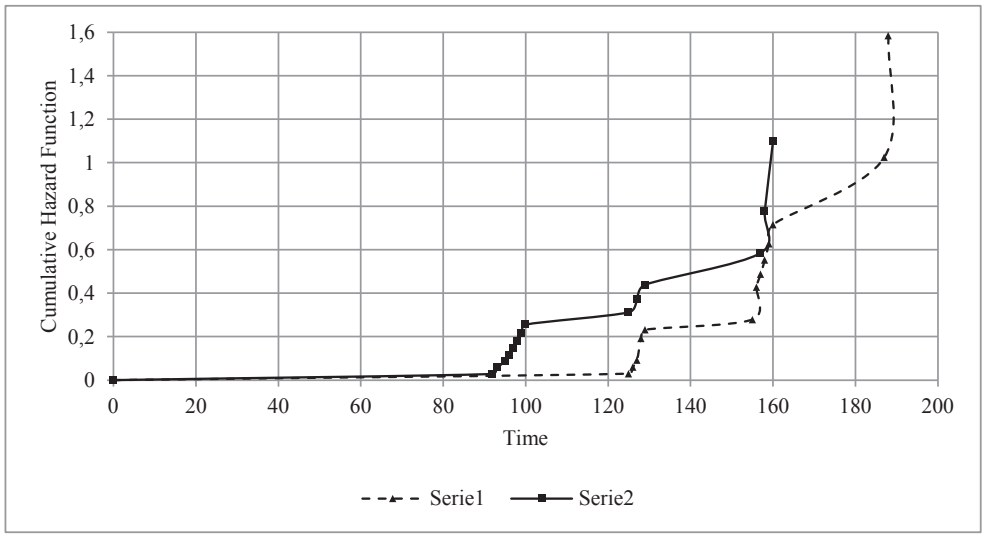

Figure 6. Plot of the cumulative hazard function by SITE

Source: own calculations by means of SAS® Base 9.3.

On the basis of the aforementioned models, it can be stated that accounting for the non-proportional hazards (if exist) provides a more detailed interpretation.

Not only is it possible to state which group is more likely to experience the event, but it is also possible to analyze the hazard ratio that is changing over time and corresponds to the varying effect of a given covariate. At the same time, it is hard to define a general rule saying which of the two methods should be chosen. On the one hand, the stratification model is easier to implement and requires less computational resources [Allison 1995]. On the other hand, if an analysis is performed with the use of a relatively small dataset or with the use of a powerful computer, resources are not that important and interaction with time might be introduced to the model. The latter approach enables obtaining parameter estimate for the covariate for which the proportional hazard assumption is violated, as well as analyze how the hazard ratio changes over time, which is impossible if the stratification model is chosen. While trying to compare two models accounting for non-proportional hazards presented earlier, i.e. Model 2 including interaction of SITE by TIME and Model 4 based on stratification, information criteria and partial likelihood function values might be compared. Statistics corresponding to each model are presented in Table 5. Additionally, fit statistics for initial model including SITE and AGE as covariates are included.

Table 5. Fit statistics comparison between models 1, 2 and 3

	SITE by TIME interaction	SITE as stratification variable	Initial model
−2lnL	163.593	142.361	168.051
AIC	169.593	144.361	172.051
SBC	173.895	145.795	174.919

Source: own calculations by means of SAS® Base 9.3.

It can be noticed that the initial model (Model 1 including SITE and AGE as covariates) which neglects the fact that the proportional hazard assumption is violated for SITE variable has the highest values of all three fit statistics. While comparing Model 2 and Model 4, it can be stated that all criteria have lower values for the stratification model, which stands for this approach. The likelihood ratio test, however, cannot be performed in this case as considered models are not nested. Additionally – in order to perform an overall assessment of both models – a linear predictor is calculated and ten binary variables are created on the basis of its percentiles. Nine out of ten binary variables are introduced to the models and their statistical significance is verified. The piece of SAS code presented as SAS Code 7 in Appendix is used to perform these procedures. In Table 6 only these newly added binary variables are included for which the parameter estimate was possible to obtain.

In the stratification model none of newly added variables is statistically significant at any acceptable level. When it comes to model with TIME by SITE interaction, two variables for which parameters estimates were obtained are significant which

Table 6. Goodness of fit comparison between models 1, 2 and 3

Model 2 (including SITE by TIME Interaction)					
Variable	Parameter Estimate	SE	Chi-square	*p*-value	Hazard Ratio
AGE	0.37049	0.10046	13.6010	0.0002	1.448
SITE	−6.95164	3.09349	5.0498	0.0246	0.001
SITE*TIME	0.04683	0.02330	4.0395	0.0444	1.048
X10	4.12871	0.02330	8.6726	0.0032	62.098
X90	−1.55707	0.67801	5.2741	0.0216	0.211
Model 4 (SITE as stratification variable)					
Variable	Parameter Estimate	SE	Chi-square	*p*-value	Hazard Ratio
AGE	0.13433	0.09317	2.0788	0.1494	1.444
X10	−15.57684	1676	0.0001	0.9926	0.000
X90	0.25073	0.59950	0.1749	0.6758	1.285

Source: own calculations by means of SAS® Base 9.3.

suggests poor fit of the model. Thus, the comparison of the information criteria as well as the linear predictor method would suggest using the stratification model rather than introducing interaction with time to the model. It should be noticed, however, that these tests should not be treated as an oracle as they give a kind of suggestions rather than unequivocal determinant of a decision which model should be chosen.

4. Conclusions

Although the proportional hazard assumption is one of the most important features in Cox model, its violation should not definitely prevent from using this statistical tool. The current paper presents two methods that were developed in order to take into account the effect of a covariate that varies through time. The performed calculations give the evidence that stratification or including interaction with time results in a better model fit in the case of covariates for which the proportional hazard assumption is not satisfied. Additionally, more detailed results and interpretation are obtained with the use of the presented method. It would be hard, however, to define a general rule for non-proportional hazards handling. An analyst can consider one of three possibilities, i.e. keeping all covariates in the model and neglecting the fact of the violation from the non-proportional hazard assumption, introducing the interaction of TIME by SITE and the estimation of the stratification model. Each of these approaches has its pros and cons. What is more, it is hard to compare these models, especially the stratification model, with non-stratified models as they differ in their construction. Thus, as soon as non-proportional hazards are identified in the model, this fact should be definitely taken into consideration, but the choice of the method needs to be adjusted for particular example data.

References

Allison P.D., 1995, *Survival Analysis Using SAS®. A Practical Guide*, SAS Institute Inc., Cary NC.
Cox D.R., 1972, Regression models and life-tables, *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187–220.
Hosmer D., Lemeshow S., 1999, *Applied Survival Analysis. Regression Modeling Time to Event Data*, John Wiley & Sons Inc., New York.
Kalbfleisch J.D., Prentice R.L., 1980, *The Statistical Analysis of Failure Time Data*, John Wiley & Sons Inc., New Jersey.
Quantin C, Moreau, T., Asselain B., Maccario J., Lellouch, J., 1996, A regression survival model for testing the proportional hazards hypothesis, *Biometrics*, no. 52, pp. 874–885.

ROZSZERZENIA MODELU COXA DLA NIEPROPORCJONALNYCH HAZARDÓW

Streszczenie: Model proporcjonalnych hazardów jest jedną z najczęściej wykorzystywanych metod w analizie czasu przeżycia. Zmienną zależną jest funkcja hazardu, która jest wyjaśniana przez czynnik zależny tylko od czasu (nazywany hazardem bazowym) oraz czynnik związany ze zmiennymi objaśniającymi. Model oparty jest na kilku restrykcyjnych założeniach. Jedno z nich dotyczy stałości ilorazu hazardów w czasie. Niemniej jednak odchylenie od tego założenia nie jest jednoznaczne z koniecznością rezygnacji ze stosowania modelu Coxa. Niniejszy artykuł prezentuje dwie metody modyfikacji modelu dla przypadku nieproporcjonalnych hazardów: wprowadzenie do modelu interakcji wybranych zmiennych objaśniających ze zmienną czasową oraz model warstwowy. Przeprowadzone kalkulacje dostarczają dowodów na to, że przedstawione metody poprawiają statystyki dopasowania modelu w porównaniu z oryginalnym modelem Coxa. Ponadto możliwa jest bardziej precyzyjna interpretacja wyników, umożliwiająca uwzględnienie zmiennego w czasie wpływu danej zmiennej objaśniającej na poziom hazardu. Wybór jednej z dwóch przedstawionych metod modyfikacji modelu Coxa dla przypadku nieproporcjonalnych hazardów nie jest jednoznaczny i powinien być dostosowany do konkretnej analizy.

Słowa kluczowe: model Coxa, analiza przeżycia, nieproporcjonalny hazard.

Appendix: SAS Codes

```
/*SAS Code 1:
Initial Cox model estimation – site and age as covariates*/
proc phreg data = a.site;
      format site site.;
      model time*censor(0) = age site / ties = exact;
run;

/*SAS Code 2:
Lifetables estimation according to Kaplan-Meier formula;
generation of the plot of 'log-negative-log' of survival func-
tion
separately for each site -> STRATA statement*/
proc lifetest data = a.site plots = (s, lls);
      format site site.;
      strata site;
      time time*censor(0);
run;

/*SAS Code 3:
Cox model estimation – age and site as covariates;
saving Schoenfeld residuals in output dataset*/
proc phreg data = a.site;
      format site site.;
      model time*censor(0) = age site / ties = exact;
      output out = schoen ressch = age_s site_s ;
run;
/*Generation of plot of Schoenfeld residuals for site as
function of time*/
proc gplot data = schoen;
      symbol1 v = dot c = black width = 1 i = sm80s;
      plot site_s*time / haxis = axis1 vaxis = axis2;
      axis1 label = ('Time');
      axis2 label = (a = 90 'Schoenfeld Residual for Site');
run;
/*SAS Code 4:
Cox model estimation – age and site as covariates, with in-
teraction of site and time added*/
proc phreg data = a.site;
      format site site.;
      model time*censor(0) = age site site_t/ ties = exact;
      site_t = site*time;
      test: test site_t;
run;
```

```
/*SAS Code 5:
Cox model estimation – AGE as a covariate, SITE as stratifica-
tion variable;
saving estimates of survival and cumulative hazard functions
in BASE dataset*/
proc phreg data = a.site;
      baseline out = base survival = surv cumhaz = cumhaz;
      format site site.;
      strata site;
      model time*censor(0) = age / ties = exact;
run;

/*SAS Code 6:
Plots of covariates-adjusted cumulative hazard and survival
functions obtained on the basis of the stratified model*/
proc gplot data = base;
      symbol1 v = star c = black width = 1 i = sm50s;
      symbol2 v = dot c = black width = 1 i = sm50s;
      plot cumhaz*time = site / haxis = axis1 vaxis = axis2;
      axis1 label = ('Time');
      axis2 label = (a = 90 'Cumulative Hazard Function');

run;

/*SAS Code 7:
Overall assessment - model with site by time interaction*/
data over1;
      set a.site;
      xbeta = 0.23313*age - 5.90164*site + 0.03985*site*time;
      count = 1;
run;
proc univariate data = over1 noprint;
      var xbeta;
      output out = perc pctlpre = p pctlpts = 10 20 30 40 50
60 70 80 90 100;
run;

data perc;
      set perc;
      count = 1;
run;

data over1a;
      merge over1 perc;
      by count;
```

```
run;

%macro ret;
data over1a;
 set over1a;

 if xbeta<=p10 then x10 = 1;
 else x10 = 0;
 if xbeta<p90 then x100 = 1;
 else x100 = 0;

%do i = 10 %to 80 %by 10;
%do j = 20 %to 90 %by 10;
      if xbeta>p&i and xbeta<=p&j then x&j = 1;
      else x&j = 0;
      %end;
      %end;
run;
%mend;

%ret;

proc phreg data = over1a;
model time*censor(0) = age site site_t x10 x20 x30 x40 x50
x60 x70 x80 x90 / ties = exact;
      site_t = site*time;
```