

Assessment of Cox Proportional Hazard Model Adequacy Using PROC PHREG and PROC GLOT

Jadwiga Borucka, Quanticate, Warsaw, Poland

ABSTRACT

Survival models are very useful in clinical trials as they are designed to perform 'time to event' analyses on data with censored observations (defined as observations with incomplete information in case subject did not experience the event during the study). Cox proportional hazard model treats hazard, i.e. conditional probability of the event occurrence under the condition that the event has not occurred so far, as a dependent variable and enables estimation of the influence of chosen covariates on the hazard level. Model estimation can be performed with one simple proc step in SAS®, however assessment of model adequacy is a little more complicated and time consuming. As a result, it is often limited to examining statistical significance of covariates. This paper describes crucial model assumptions and presents the process of model assessment with the use of SAS software.

INTRODUCTION

• Introduction to Survival Analysis

Survival models are being applied in studies where the time to occurrence of a defined event is analyzed. While performing survival analysis one has to precisely define for each subject:

- **beginning of the time period during which a subject is being observed**, which is a moment when a subject joined the study,
- **end of the time period during which a subject is being observed**, which is: a moment of the event occurrence (for subjects that experienced the event) or a moment of withdrawal from study (for subjects that were withdrawn for any reason) or the end of the study (for subjects that completed the study not experiencing the event),
- **indicator whether a subject experienced the event**,
- **time variable** - survival time that is equal to the length of time period during which a subject is being observed.

What distinguishes survival analysis from other regression models is the fact that for subjects that completed the study without experiencing the event, the time period during which they are observed is truncated (censored) at the end of the study or at the moment of withdrawal.

The following functions are applied in survival analysis:

- **Cumulative Density Function** - probability that survival time will not be greater than the given value t :

$$F(t) = P(T \leq t) = \int_0^t f(z)dz$$

- **Density Function** - a change of probability of the event occurrence during an infinitesimal time period of the length Δt ; a derivative of cumulative density function:

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t} = F'(t)$$

- **Survival Function** - probability that a subject will survive (i.e. will not experience the event) at least to the moment t :

$$S(t) = P(T \geq t) = 1 - F(t)$$

- **Hazard Function** - conditional probability of the event occurrence at the moment t under condition that a subject has not experienced the event so far:

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}$$

- **Cumulative Hazard Function** - integral of the hazard function which is interpreted as the expected number of events to the moment t :

$$H(t) = \int_0^t \lambda(z) dz = \int_0^t \frac{f(z)}{1 - F(z)} dz = -\ln(1 - F(z))|_0^t = -\ln(1 - F(t)) = -\ln S(t)$$

A general assumption of survival models that are designed to analyze influence of subject specific characteristics on the survival time is that the hazard can be calculated as a product of two functions. The first one characterizes how the hazard changes as a function of survival time (baseline hazard function), the second - how it changes as a function of covariates. Thus, a general formula for hazard function, which in fact is a function of both survival time and covariates, is as follows:

$$\lambda(t, \mathbf{z}) = \lambda_0(t) \cdot \lambda_1(\mathbf{z})$$

where $\lambda(t, \mathbf{z})$ is the hazard function, $\lambda_0(t)$ is a component related to survival time and $\lambda_1(\mathbf{z})$ is a component related to covariates vector $\mathbf{z}$.

Depending on the specification of the above mentioned functions, survival models can be divided into three categories: non - parametric models that estimate hazard function without assuming any specific mathematical formula for the components, semi - parametric models that define precisely the covariates - related component and parametric models that parameterize both components.

• Cox Proportional Hazard Model Definition

Cox proportional hazard model is one of the semiparametric regression models applied in survival analysis. It uses exponential function to define covariates - related component of the hazard function which results in the following formula of the model:

$$\lambda(t, \mathbf{z}) = \lambda_0(t) \cdot \exp(\beta \mathbf{z})$$

where β is a vector of parameters estimates.

Having defined hazard function, one can derive survival function which is expressed as:

$$S(t, \mathbf{z}) = \exp[-H(t, \mathbf{z})]$$

where $H(t, \mathbf{z})$ is a cumulative hazard function defined as:

$$H(t, \mathbf{z}) = \int_0^t \lambda(s, \mathbf{z}) ds = \int_0^t \lambda_0(s) \exp(\beta \mathbf{z}) ds = \exp(\beta \mathbf{z}) \int_0^t \lambda_0(s) ds = \exp(\beta \mathbf{z}) H_0(t)$$

where $H_0(t)$ is cumulative baseline hazard function. As a result, survival function is expressed as:

$$S(t, \mathbf{z}) = \exp[-\exp(\beta \mathbf{z}) H_0(t)] = S_0(t)^{\exp(\beta \mathbf{z})}$$

where $S_0(t)$ is a baseline survival function.

• Model Assumptions

There are four crucial assumptions that are being made while Cox proportional hazard model estimation and assessment:

- **Covariates have multiplicative effect on the hazard function** which results directly from the model formula¹; this assumption is not being formally examined,
- **Covariates have linear effect on natural logarithm of the hazard function** which results from the model formula that can be transformed to the following equation by using the natural logarithm:

$$\ln \lambda(t) = \ln \lambda_0(t) + \beta \mathbf{z}$$

¹The formula in which covariates have multiplicative effect on hazard scale is characteristic for proportional hazard models and distinguishes them from accelerated time failure models, in which covariates have multiplicative effect on the time scale.

- **Hazard ratio for the two subjects is constant over survival time** (proportional hazard assumption) which results from the hazard ratio calculation:

$$HR = \frac{\lambda(t, \mathbf{z}_1)}{\lambda(t, \mathbf{z}_2)} = \frac{\lambda_0 \exp(\beta \mathbf{z}_1)}{\lambda_0 \exp(\beta \mathbf{z}_2)} = \exp(\mathbf{z}_1 - \mathbf{z}_2)' \beta$$

It can be easily noticed that the hazard ratio (HR) does not depend on the time variable value but only on covariates' values.

- **There are no tied events**, i.e. events that occur exactly at the same time. Ideally, time variable is assumed to be continuous. In fact, time is usually measured in minutes, hours, days, etc., thus it is not unusual to observe two events happening at the same time. However, software packages are currently developed so that they could handle such cases properly and violations from this assumption do not cause any serious problems anymore².

In practice, the second and the third assumption are being examined in proportional hazard model adequacy assessment process which is described in the next sections.

RESIDUALS

Cox proportional hazard model differs from classical regression models due to the time variable construction which is truncated for subjects not experiencing the event. As a result, it is impossible to calculate a difference between the actual (observed) and expected (resulting from the model) value of dependent variable which is a well - known definition of a model residual. Semiparametric survival models led to developing other types of residuals that are being used in the model adequacy assessment process. There are three main types of residuals defined for Cox proportional hazard model:

- **Martingale Residuals**

As far as one - event³ models are concerned, martingale residual for the i^{th} subject at the moment t is defined as follows:

$$M_i(t) = \sigma_i(t) - H(t, \mathbf{z}_i)$$

and is interpreted as a difference between the actual and expected number of event occurrence till the moment t . Usually martingale residuals are subject - specific and are calculated as of at the end of the study. As residuals of this type do not have symmetric distribution, they can be transformed into deviance residuals that are supposed to have a symmetric distribution with the mean equal to zero, assuming proper specification of the model. Martingale residuals are useful while examining assumption of linear effect of covariates on logarithm of hazard, while deviance residuals help also in identifying poorly fitted subjects, which is described in detail in the section dedicated to assumptions examining.

- **Score Residuals**

Score residuals are subject - and covariate - specific (i.e. they can be calculated for each subject with respect to each covariate) and are defined as follows:

$$\frac{\partial L(\beta)}{\partial \beta_k} = \sum_{i=1}^n L_{ik}$$

where L_{ik} is so called score process for the i^{th} subject with respect to the k^{th} covariate and is expressed as:

$$L_{ik} = \sum_{j=1}^n (\mathbf{z}_{ik} - \bar{\mathbf{z}}_{w_{jk}}) dM_i(t_j)$$

where z_{ik} is a value of the k^{th} covariate for the i^{th} subject and:

²The most precise method of tied events handling was developed by Kalbfleish and Prentice. Their idea was to consider all possible combinations of order in which tied events occur. This method can be obtained by using EXACT option in MODEL statement in PHREG procedure.

³One - event models assume that only one event can occur for each subjects, i.e. in case when more than one event if fact occur, only the first one is taken into consideration.

$$\bar{\mathbf{z}}_{w_j k} = \frac{\sum_{l \in R(t_j)} \mathbf{z}_{lk} \exp(\beta \mathbf{z}_l)}{\sum_{l \in R(t_j)} \exp(\beta \mathbf{z}_l)}$$

Expression $dM_i(t_j)$ is a change of martingale residual for the i^{th} subject at the moment t_j , defined as:

$$dM_i(t_j) = dN_i(t_j) - Y_i(t_j) \exp(\beta \mathbf{z}_i) \lambda_0(t_j)$$

where $\lambda_0(t_j)$ is value of baseline hazard at the moment t_j . Expression $dN_i(t_j)$ is a change of value of function that is a sum of events for the i^{th} subject to the moment t_j . In one - event studies the function is constantly equal to zero for subjects not experiencing the event, for subjects that experience the event it is equal to zero until the event occurrence and one afterwards. $Y(t_j)$ is risk process defined as:

$$Y_i(t_j) = 1 \text{ if } t_i \geq t_j \text{ and } Y_i(t_j) = 0 \text{ otherwise.}$$

Score residual for the i^{th} subject with respect to the k^{th} covariate can be interpreted then as a weighted difference between the value of the covariate for the given subject and a weighted average value of the covariate in a risk set⁴, where the weight is a change of martingale residual for the given subject at the given time point. Score residuals are useful mainly while identifying outliers and influential subjects in a sample. In the latter situation scaled score residuals, called also dfbeta residuals, can be helpful as well. Dfbeta residuals are calculated by dividing score residuals by variance of parameter estimate for the given covariate. Scaled score residual for the i^{th} subject with respect to the k^{th} covariate is interpreted as approximate change in parameter estimate for the k^{th} covariate after exclusion of the i^{th} subject from the population.

• Schoenfeld Residuals

The last type of residuals that are discussed are so called Schoenfeld residuals. They express 'input' of the i^{th} subject in the derivative of logarithm of partial likelihood function⁵ with respect to the given covariate⁶ and are calculated only at moments of event occurrence. Mathematical formula for Schoenfeld residual is as follows:

$$\frac{\partial L_p(\beta)}{\partial \beta_k} = \sum_{i=1}^n \sigma_i \mathbf{z}_{ik} - \frac{\sum_{j \in R(t_i)} \mathbf{z}_{jk} \exp(\beta \mathbf{z}_j)}{\sum_{j \in R(t_i)} \exp(\beta \mathbf{z}_j)} = \sum_{i=1}^n \mathbf{z}_{ik} - \bar{\mathbf{z}}_{w_{ik}} \text{ gdzie: } \bar{\mathbf{z}}_{w_{ik}} = \frac{\sum_{j \in R(t_i)} \mathbf{z}_{jk} \exp(\beta \mathbf{z}_j)}{\sum_{j \in R(t_i)} \exp(\beta \mathbf{z}_j)}$$

Schoenfeld residual for the k^{th} covariate and the i^{th} subject can be calculated as:

$$r_{ik} = \sigma_i(\mathbf{z}_{ik})$$

Residuals of this type are used mostly while examining assumption of proportional hazards.

ASSESSMENT OF MODEL ADEQUACY

Assessment of survival model adequacy can be divided into five steps. The first step is verification of statistical significance of covariates which can be done in a similar way as in case of other regression models. Then, Cox proportional hazard model specific assumptions are being examined, namely: linear effect of covariates on the logarithm of hazard and proportional hazard assumption. In the next step, poorly fitted and influential subjects are being identified. Finally, an overall goodness - of - fit test is being performed. In the present subsection assessment of model adequacy will be presented using a theoretical sample of 100 subjects, 14 of whom experienced the event. There are two covariates: gender (binary variables equal to 1 for males and 0 for females) and age (continuous variable).

⁴As a risk set at the moment t one understands the population of subjects that have not experienced the event till the given time point t .

⁵Partial likelihood function is a likelihood function used in model estimation, modified for the proportional hazard model estimation purpose so that it took into consideration truncated values of time variable.

⁶Schoenfeld residual for the i^{th} subject with respect to the k^{th} covariate can be also interpreted as a difference between actual value of the k^{th} covariate for the i^{th} subject and expected value of the k^{th} covariate in a risk set at the given time point.

- **Statistical Significance of Covariates**

While examining statistical significance of covariates three tests are being used: partial likelihood ratio test, score test and Wald test. Null hypothesis is always the same and assumes that parameters for all covariates are equal to 0. The first test is based on the partial likelihood function that is used in estimation process. Test statistics is expressed as follows:

$$\chi_{LR}^2 = 2[\ln L(\hat{\beta}) - \ln L(0)]$$

where $\ln L(0)$ is a logarithm of partial likelihood function from model with no covariates and $\ln L(\hat{\beta})$ is a logarithm of partial likelihood function from a model of interest. The statistics has chi - square distribution with p degrees of freedom, where p is a number of covariates in a model.

The Wald test uses the statistics:

$$\chi_W^2 = \hat{\beta}'[\hat{V}(\hat{\beta})]^{-1}\hat{\beta}$$

where $V(\beta)$ is a variance - covariance matrix of parameter estimates. The statistics has a standard normal distribution.

The score test is based on the following statistics:

$$\chi_s^2 = \left[\frac{\partial L(0)}{\partial \beta}\right]' I^{-1}(0) \left[\frac{\partial L(0)}{\partial \beta}\right]$$

where $I(0)$ is information matrix defined as: $I(0) = -\frac{\partial^2 L(0)}{\partial \beta^2}$. The test statistics has a standard normal distribution.

The following SAS code enables to obtain results of above described tests of statistical significance of covariates:

```
proc phreg data = sample;
    model time*censor(0) = age gender / ties = exact;
run;
```

which gives the output:

Testing Global Null Hypothesis: BETA=0							
		Test	Chi-Square	DF	Pr > ChiSq		
		Likelihood Ratio	27.0927	2	<.0001		
		Score	62.3108	2	<.0001		
		Wald	30.6589	2	<.0001		
The PHREG Procedure							
Analysis of Maximum Likelihood Estimates							
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Variable Label
AGE	1	-0.11147	0.04777	5.4442	0.0196	0.895	AGE
GENDER	1	1.87843	0.81161	5.3566	0.0206	6.543	GENDER

As it can be read from the first part of the output, all tests of joint statistical significance are coherent and lead to conclusion that all variables are statistically significant. Considering each of the two covariates separately, the conclusion is similar - each of covariates is significant at the level 0.05.

- **Linear Relation between Covariates and Logarithm of Hazard**

The first of the two Cox proportional hazard model assumptions that are examined in practice can be verified in the two ways. The first method reestimates the model omitting the variable of interest and calculates martingale residuals. A plot of those residuals versus the analyzed variable reflects the type of relation that binds the given covariate and logarithm of hazard. Thus, if residuals form a straight line it can be stated that the assumption is satisfied. The second method is applicable only for continuous variables. A variable of interest is being categorized, basing for example on quantiles calculated from the

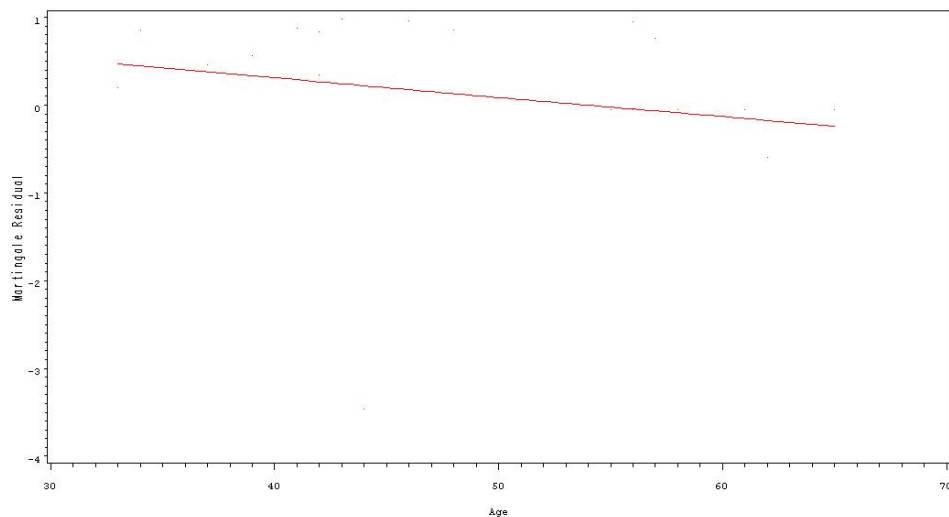
covariate distribution. Then, for each interval, a corresponding binary variable is being created which is equal to 1 if a variable value falls within the given interval and 0 otherwise. Such binary variables are included in the model and estimates of parameters for additional binary variables are being analyzed. The best way is a graphical analysis, i.e. drawing a plot of parameters' estimates versus centers of earlier defined intervals - a straight line suggests that the assumption is satisfied.

The following SAS code can be used to implement the first method:

```
proc phreg data = sample;
    model time*censor(0) = gender / ties = exact;
    output out = martingale resmart = resmart;
    id age;
run;

proc gplot data = martingale;
    plot resmart*age / haxis = axis1 vaxis = axis2;
    symbol v = point c = red width = 1 i = sm90s;
    axis1 label = ('Age');
    axis2 label = (a = 90 'Martingale Residual');
run;
```

producing the output:



In order to get the results from the second method, the following code comes in handy:

```
data sample; set sample;
    if age<=40 then w1 = 1; else w1 = 0;
    if age>40 and age<=50 then w2 = 1; else w2 = 0;
    if age>50 and age<=60 then w3 = 1; else w3 = 0;
    if age>60 then w4 = 1; else w4 = 0;
run;

proc phreg data = sample outest = loglinear;
    model time*censor(0) = gender w1 w2 w3 /ties = exact;
run;

proc transpose data = loglinear out = age;
    var w1 w2 w3;
run;

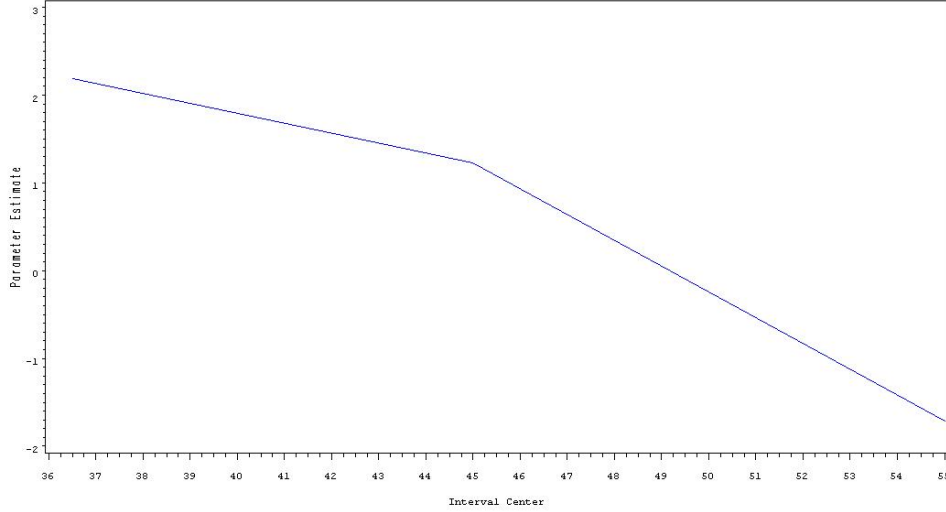
data age; set age;
    if _name_ = 'w1' then point = 36.5;
    if _name_ = 'w2' then point = 45;
    if _name_ = 'w3' then point = 55;
run;
```

```

proc gplot data = age;
  plot time*point / haxis = axis1 vaxis = axis2;
  symbol 1 = join c = blue line = 1;
  axis1 label = ('Interval Center');
  axis2 label = (a = 90 'Parameter Estimate');
run;

```

Code produces the following plot:



Both methods lead to the similar conclusion - age has linear effect on logarithm of hazard. Martingale residuals form a straight line, parameter estimates for created binary variables similarly. Thus, the assumption of log - linear relation between hazard and age is satisfied.

- **Identification of Influential and Poorly Fitted Subjects**

Outliers in Cox proportional hazard model can be divided into three categories. There are: subjects that have value of a covariate that differs from the sample average to a great extent (1), subjects that have strong influence on parameters estimates (2) and subjects that have strong influence on the partial likelihood function value and thus on the model adequacy (3). First set of subjects is being identified with the use of score residuals. After model estimation, score residuals for each subject with respect to a covariate of interest are calculated and a plot presenting values of residuals versus the analyzed covariate is drawn. Outliers are usually easily identified during such an analysis. The next step after identification of outliers is estimation of the strength of influence that suspected subjects have on parameters' estimates. A suspected subject is excluded from the sample and the model is reestimated. A change in parameter estimate is calculated as:

$$\Delta\beta_{ki} = \beta_k - \beta_{k(-i)}$$

where β_k is the estimate of parameter for k^{th} covariate in model estimated on the whole sample, while $\beta_{k(-i)}$ is a corresponding value in a model on a limited sample. It was proved that Δ_i is well approximated with the k^{th} element of the vector defined as:

$$\Delta\beta_i = Var(\beta)L_i$$

where L_i is a vector of score residuals for the i^{th} subject and $Var(\beta)$ - a variance - covariance matrix of parameters estimates. This expression is called scaled score residuals or dfbeta residuals. A plot of residuals of this type versus the analyzed covariate is helpful in identifying influential subjects. In order to estimate the influence of the given subject on the value of the partial likelihood function, the following statistics that uses score residuals is constructed:

$$ld_i = \Delta\beta'Var(\beta_i)^{-1}\Delta\beta_i = L_i'Var(\beta)Var(\beta)^{-1}Var(\beta)L_i = L_i'Var(\beta)L_i$$

The above statistics is called likelihood displacement and approximates a change in logarithm of partial likelihood function after exclusion of the i^{th} subject from the sample. Additionally, for the matrix

$L'_i Var \beta L_i$ eigenvalues are being found. Eigenvector connected with the highest eigenvalue is called 1 - max statistics. Both likelihood displacement and 1 - max statistics are plotted versus a summary statistics (for example martingale residuals). Highly influential subjects are easily identified on such a plot. SAS code that estimates the model, saving all necessary residuals:

```
proc phreg data = sample;
  model time*censor(0) = gender age / ties = exact;
  output out = score
  ressko = sc_gen sc_age
  dfbeta = df_gen df_age
  lmax = lmax
  ld = ld
  resmart = resmart;
  id obs;
run;
```

Below, there are codes that enable to obtain desired plots:

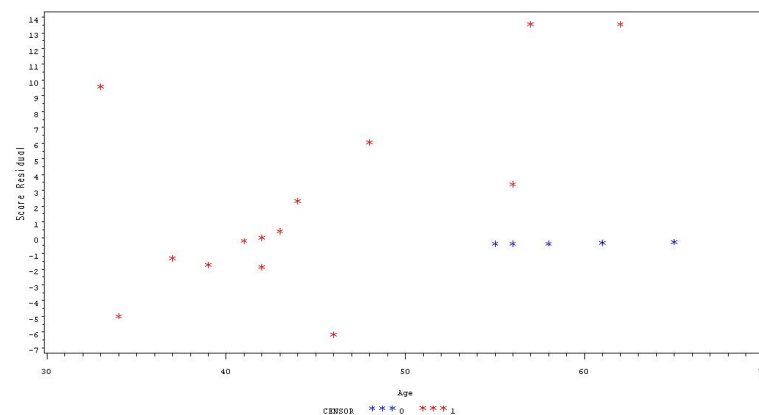
```
proc gplot data = score;
  plot sc_age*age = censor / haxis = axis1 vaxis = axis2;
  symbol1 c = blue width = 2 v = star i = none;
  symbol2 c = red width = 2 v = star i = none;
  axis1 label = ('Age');
  axis2 label = (a = 90 'Score Residual');
run;

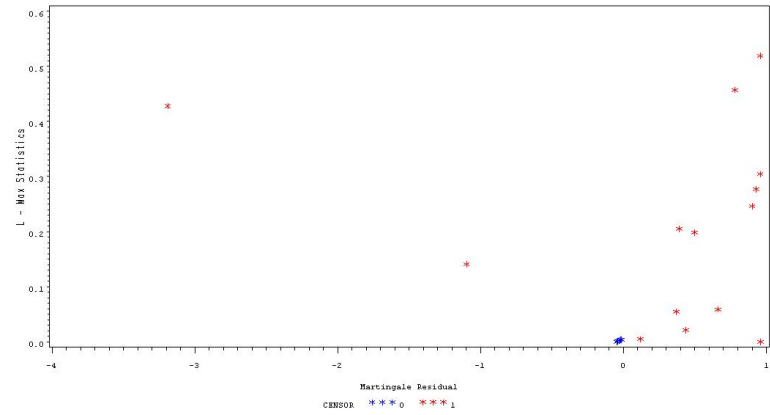
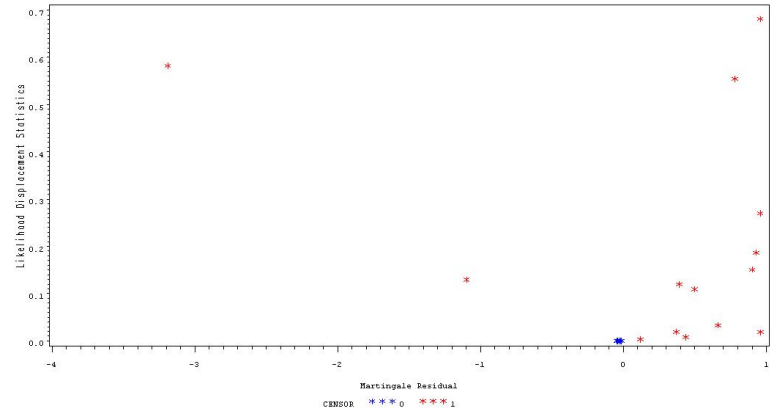
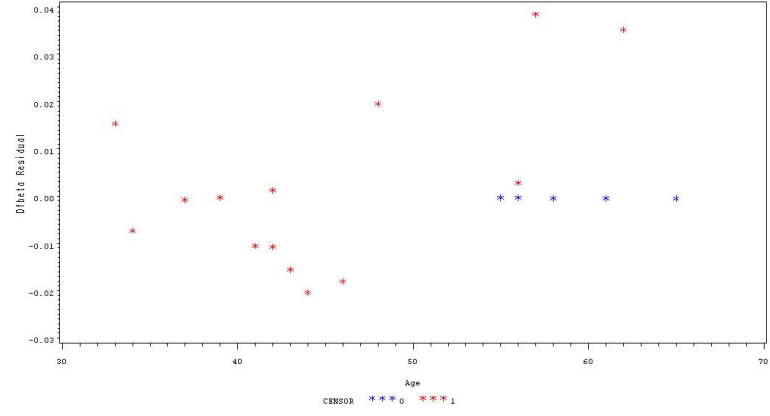
proc gplot data = score;
  plot df_age*age = censor / haxis = axis1 vaxis = axis2;
  symbol1 c = blue width = 2 v = star i = none;
  symbol2 c = red width = 2 v = star i = none;
  axis1 label = ('Age');
  axis2 label = (a = 90 'Dfbeta Residual');
run;

proc gplot data = score;
  plot lmax*resmart = censor / haxis = axis1 vaxis = axis2;
  symbol1 c = blue width = 2 v = star i = none;
  symbol2 c = red width = 2 v = star i = none;
  axis1 label = ('Martingale Residual');
  axis2 label = (a = 90 'L - Max Statistics');
run;

proc gplot data = score;
  plot ld*resmart = censor / haxis = axis1 vaxis = axis2;
  symbol1 c = blue width = 2 v = star i = none;
  symbol2 c = red width = 2 v = star i = none;
  axis1 label = ('Martingale Residual');
  axis2 label = (a = 90 'Likelihood Displacement Statistics');
run;
```

Above codes produced plots:





There are three subjects on the score residuals plot that have visibly higher value of score residual with respect to age than other subjects in the sample, they can be easily distinguished also on the plot of dfbeta residuals. Both plots: likelihood displacement and l - max statistics versus martingale residuals indicate these three observations even more clearly. As the next step these observations should be identified and it should be verified whether excluding them from the sample could improve overall goodness of fit of the model (by comparing values of information criteria and using likelihood ratio test). The current paper however is prepared only for presentation the purpose so this step will be skipped at the moment.

- **Proportional Hazard Assumption**

There are two methods of checking the proportional hazard assumption. The first one adds time - dependent variables to the original model (in this case, product of a variable of interest and logarithm of time variable was added for each covariate). If such variable is statistically significant then it can be concluded that the assumption of proportional hazards is not satisfied for the given covariate. The other method draws plot of Shoenfeld residuals versus time variable. If the obtained line is approximately horizontal

then it can be concluded that proportional hazard assumption is satisfied.

The following code estimates model with time - dependent variables:

```
proc phreg data = sample;
  model time*censor(0) = gender age g_time a_time / ties = exact;
  g_time = gender*log(time);
  a_time = age*log(time);
run;
```

Model estimation results:

The PHREG Procedure							
Analysis of Maximum Likelihood Estimates							
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Variable Label
GENDER	1	10.76654	4.74708	5.1440	0.0233	47407.82	GENDER
AGE	1	0.03012	0.19739	0.0233	0.8787	1.031	AGE
g_time	1	-2.58024	1.33373	3.7427	0.0530	0.076	
a_time	1	-0.03118	0.06871	0.2059	0.6500	0.969	

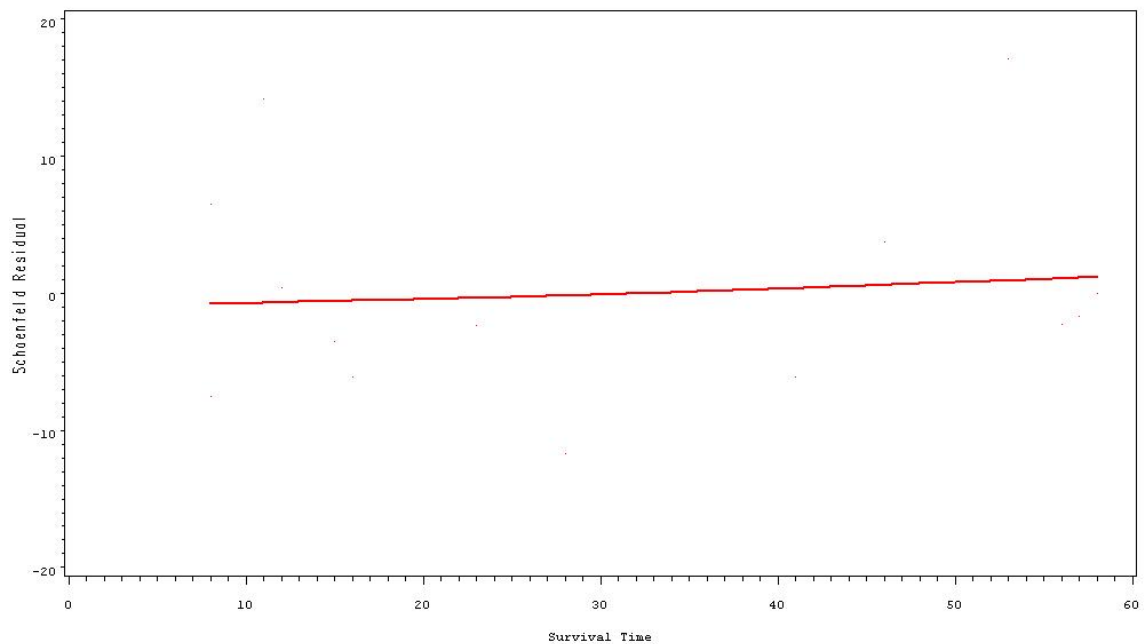
None of the time - dependent variables that were added to the model is statistically significant which indicates that proportional hazard assumption for both covariates is satisfied.

The following code enables to estimate the model and to save Schoenfeld residuals. In the next step a plot of Schoenfeld residuals with respect to age versus survival time is being drawn.

```
proc phreg data = sample;
  model time*censor(0) = gender age / ties = exact;
  output out = schoen
  ressch = sc_gen sc_age;
run;

proc gplot data = schoen;
  plot sc_age*time = 1 / haxis = axis1 vaxis = axis2;
  symbol c = red v = point i = sm90s width = 2;
  axis1 label = ('Survival Time');
  axis2 label = (a = 90 'Schoenfeld Residual');
run;
```

The code produced the plot:



Conclusions from the above plot are consistent with results of time - dependent variables analysis and suggest that assumption of proportional hazard is satisfied as far as age is concerned.

- **Overall Assessment of the Model Adequacy**

The last step in the process of assessment of model adequacy is an overall assessment of the model. There are some well developed methods of relative model evaluation, however they are commonly used also for linear regression models and as such not discussed in the present paper (e.g. likelihood ratio test, information criteria). This section describes two ways of general assessment of model adequacy that are designed for Cox proportional hazard model. The first method uses the simplified test developed by Gronnesby and Borgan. There is a linear predictor $\beta\mathbf{z}$ for each subject calculated. Then, the sample is divided into K groups, on the basis of linear predictor values. For each group, binary variable is being created that indicates whether a subject is included in the given group or not. K - 1 binary variables are added to the model, lack of statistically significance of K - 1 additional variables indicates well fitted model. The other method employs a plot of cumulative number of actual events versus cumulative expected number of events. If the model is well fitted, points on the plot should follow a 45 degree line beginning at the origin.

The code below is used to perform Gronnesby and Borgan test:

```
data sample; set sample;
    xbeta = 1.87843 * gender - 0.11147 * age;
run;

proc univariate data = sample noprint;
    var xbeta;
    output out = xbeta
    pctlpts = 10 20 30 40 50 60 70 80 90 100
    pctlpre = xb
    pctlname = p10 p20 p30 p40 p50 p60 p70 p80 p90 p100;
run;

data sample;
    merge sample xbeta;
    licz = 1;
run;

%macro retain;

%do i = 10 %to 100 %by 10;
    data sample; set sample;
        by licz;
        retain p&i;
        if first.licz then
            p&i = xbp&i;
    run;
%end;

data sample; set sample;
    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;

%retain;

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

Results:

The PHREG Procedure						
Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
GENDER	1	0.79324	0.70054	1.2822	0.2575	2.211
AGE	1	-0.06102	0.06135	0.9893	0.3199	0.941
x10	1	-3.17656	1.65183	3.6981	0.0545	0.042
x20	0	0	.	.	.	.
x30	0	0	.	.	.	.
x40	0	0	.	.	.	.
x50	0	0	.	.	.	.
x60	0	0	.	.	.	.
x70	0	0	.	.	.	.
x80	0	0	.	.	.	.
x90	1	1.05619	1.15176	0.8409	0.3591	2.875

None of the additional binary variables is statistically significant at the alpha level equal to 0.05 which indicates a well fitted model.

The following code is used to obtain overall assessment of the model according to the second method:

```
proc phreg data = sample;
    model time*censor(0) = gender age / ties = exact;
    output out = est survival = surv logsurv = ls loglogs = ll;
run;
proc sql;
    select sum(censor) into: event
    from sample;
quit;
proc sql;
    select count(*) into: obs
    from sample;
quit;
data est; set est;
    hazard = -log(surv);
    czek = hazard + ls;
run;
proc sort data = est; by time descending surv; run;

data est; set est;
    nr = 1;
run;
data est; set est;

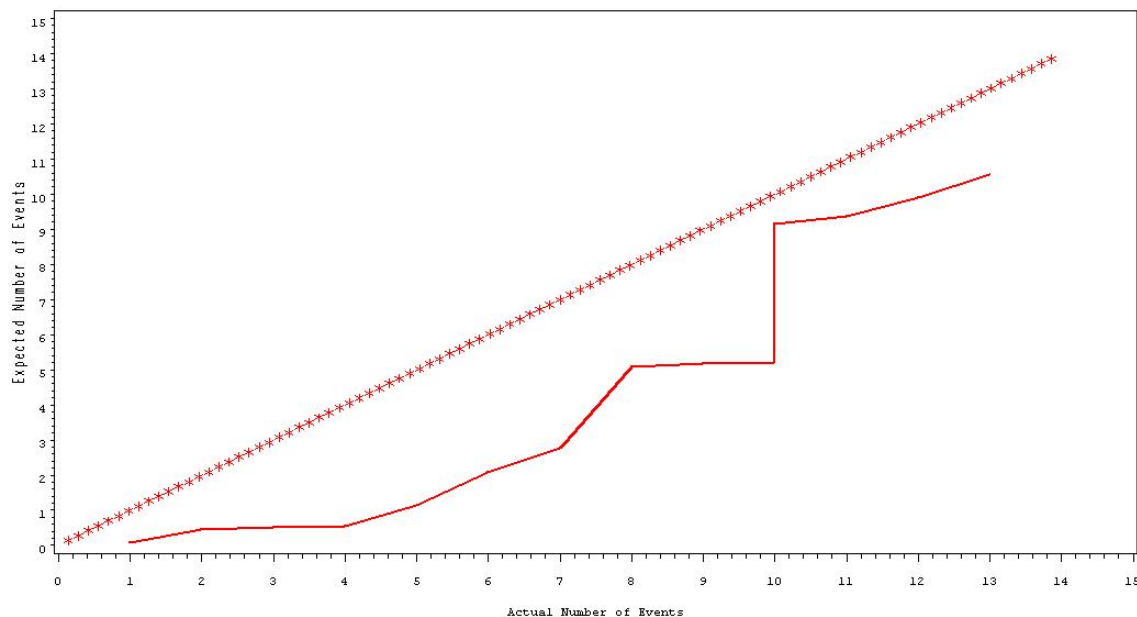
    retain event_ hazard_;
    by nr;
    if first.nr then do;
        event_ = censor;
        hazard_ = hazard;
    end;

    else do;
        event_ = event_ + censor;
        event_ = event_;
        hazard_ = hazard_ + hazard;
        hazard_ = hazard_;
    end;

    retain y 0;
    y +1;
    count = y*&event/&obs;

run;
proc gplot data = est (where = (hazard^=.));
    plot hazard_*event_ count*count / overlay haxis = axis1 vaxis = axis2;
    symbol1 v = point i = join c = blue;
    symbol1 v = point i = join c = red;
    axis1 label = ('Actual Number of Events') order = (0 to 15 by 1);
    axis2 label = (a = 90 'Expected Number of Events') order = (0 to 15 by 1);
run;
```

As a result, the following plot was obtained:



The line on the above plot has the right direction, however it differs from a 45 degree line to a great extent. It may be, among others, due to violations from model assumption for the other covariate: gender (assumptions were examined only with respect to age), three outliers that were not excluded from the sample or quite small number of observations. In case of not satisfying results of the overall tests of model adequacy, all previously examined assumptions should be reviewed again and - if necessary - model specification amended.

CONCLUSIONS

The current paper describes the process of model adequacy assessment for Cox proportional hazard model, which seems to be often neglected. It is crucial to follow the algorithm step by step. Using a model that did not go through the complex process of model adequacy assessment successfully might be risky and statistical analyzes performed on the basis of such a model can result in misleading and improper conclusions.

REFERENCES

- Arjas E., 1998, *A graphical method for assessing goodness - of - fit in Cox's proportional hazard model*, Journal of American Statistical Association.
- Cox D. R., 1972, *Regression models and life tables (with discussion)*, Journal of Royal Statistics Society.
- Gronnesby J. K., Borgan O., 1996, *A method for checking regression models in survival analysis based on the risk score*, Lifetime Data Analysis.
- Hosmer D., Lemeshow S., 1999, *Applied survival analysis. Regression modeling time to event data*, New York.
- Kalbfleisch J. D., Prentice R. L., 1980, *The statistical analysis of failure time data*, New York.
- Therneau T. M., Grambsch P. M., 2000, *Modeling survival data*, Berlin.

ACKNOWLEDGEMENTS

Special thanks to Marzena Marcinowska for time spent on reviewing this paper and important remarks.

CONTACT INFORMATION

Jadwiga Borucka, Quanticate Polska Sp. z o.o.
Hankiewiczza 2, 02 - 103 Warsaw, Poland
Tel. +48(0) 22 5762140, Fax: +48(0) 22 5762159, E - mail: jadwiga.borucka@quanticate.com