

PhUSE 2010
Paper SP05



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

Jadwiga Borucka
Quanticate, Warsaw, Poland

PRESENTATION PLAN



PRESENTATION PLAN



Brief Introduction to Survival Analysis:

- Basic definitions
- Functions used in survival analysis

PRESENTATION PLAN



Brief Introduction to Survival Analysis:

- Basic definitions
- Functions used in survival analysis

Cox Proportional Hazard Model:

- Model definition
- Residuals in Cox model

PRESENTATION PLAN



Brief Introduction to Survival Analysis:

- Basic definitions
- Functions used in survival analysis

Cox Proportional Hazard Model:

- Model definition
- Residuals in Cox model

Assessment of Model Adequacy:

- Statistical Significance of Covariates
- Linear Relation Between Covariates and Hazard
- Identification of Influential and Poorly Fitted Subjects
- Proportional Hazard Assumption
- Overall Assessment of the Model Adequacy

BRIEF INTRODUCTION TO SURVIVAL ANALYSIS



Survival models are designed to perform 'time to event' analyzes on data with censored observations (defined as observations with incomplete information in case subject did not experience the event during the study).

BRIEF INTRODUCTION TO SURVIVAL ANALYSIS



Survival models are designed to perform 'time to event' analyzes on data with censored observations (defined as observations with incomplete information in case subject did not experience the event during the study).

Each subject in a sample has to have defined:

- **beginning** of the observation period,
- **end** of the observation period,
- variable that indicates whether a subject experienced the **event**,
- **time** variable.

BRIEF INTRODUCTION TO SURVIVAL ANALYSIS



*Note: For subjects that experience the event we have complete information about the length of the period of observation, for subjects that were withdrawn from study for any reason or completed the study without experiencing the event, **time variable is censored** at the end of the study. **Analyzing of time variable that is truncated, i.e. does not reflect the actual value from the beginning of observation till the event occurrence, is characteristic for survival models.***

Subjects who experienced the event	Subjects who were withdrawn or completed the study without experiencing the event
Actual value of time variable	Censored value of time variable

BRIEF INTRODUCTION TO SURVIVAL ANALYSIS



Crucial functions in survival models:

BRIEF INTRODUCTION TO SURVIVAL ANALYSIS



Crucial functions in survival models:

➤ **Cumulative Density Function:** $F(t) = P(T \leq t) = \int_0^t f(z)dz$

BRIEF INTRODUCTION TO SURVIVAL ANALYSIS



Crucial functions in survival models:

➤ **Cumulative Density Function:** $F(t) = P(T \leq t) = \int_0^t f(z)dz$

➤ **Survival Function:** $S(t) = P(T \geq t) = 1 - F(t)$

BRIEF INTRODUCTION TO SURVIVAL ANALYSIS



Crucial functions in survival models:

➤ **Cumulative Density Function:** $F(t) = P(T \leq t) = \int_0^t f(z)dz$

➤ **Survival Function:** $S(t) = P(T \geq t) = 1 - F(t)$

➤ **Hazard Function:**

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

BRIEF INTRODUCTION TO SURVIVAL ANALYSIS



Crucial functions in survival models:

➤ **Cumulative Density Function:** $F(t) = P(T \leq t) = \int_0^t f(z)dz$

➤ **Survival Function:** $S(t) = P(T \geq t) = 1 - F(t)$

➤ **Hazard Function:**

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

➤ **Cumulative Hazard Function:**

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

COX PROPORTIONAL HAZARD MODEL



Cox Proportional Hazard Model



Hazard as a product of time – related
baseline hazard and covariates –
related component



Specific formula for covariates –
related component and undefined
baseline hazard
(semiparametric model)

Hazard as dependent variable

Model definition

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

Baseline hazard

Covariates – related component

COX PROPORTIONAL HAZARD MODEL

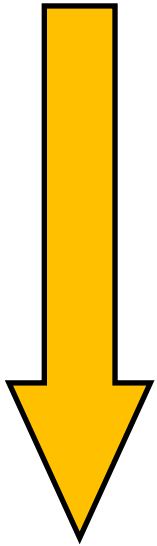


**Types of residuals calculated
for the Cox proportional hazard model**

COX PROPORTIONAL HAZARD MODEL



**Types of residuals calculated
for the Cox proportional hazard model**

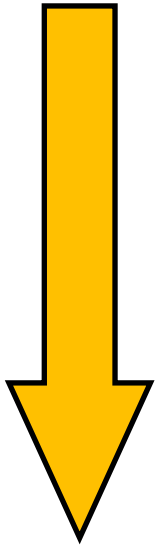


**Martingale
Residuals**

COX PROPORTIONAL HAZARD MODEL



Types of residuals calculated
for the Cox proportional hazard model



**Martingale
Residuals**

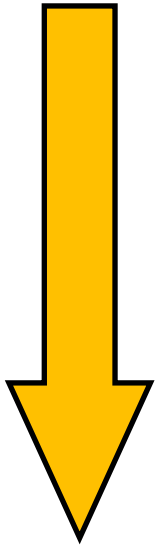


**Score
Residuals**

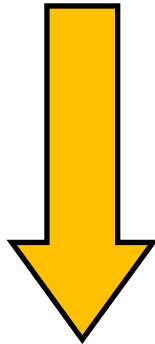
COX PROPORTIONAL HAZARD MODEL



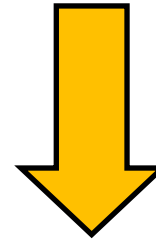
Types of residuals calculated
for the Cox proportional hazard model



**Martingale
Residuals**



**Score
Residuals**



**Schoenfeld
Residuals**

➤ Martingale Residuals

- *calculated for the given **subject**, at the given **timepoint** t ,*
- *interpreted as a **difference between actual (observed) and expected (resulting from the model) number of events** till the given timepoint t .*

➤ Score Residuals

- *calculated for the given **subject**, with respect to the given **covariate**,*
- *interpreted as a weighted **difference between value of the given covariate for the given subject and average value of this covariate in a risk set**,*
- *scaling score residuals by dividing them by the parameter estimate for the given covariate results in **dfbeta residuals** that can be interpreted as **approximate change in parameter estimate for the given covariate, after excluding from the sample particular subject**.*

➤ Schoenfeld Residuals

- *calculated for the given **subject**, with respect to the given **covariate**,*
- *interpreted as ‘**input**’ of a given subject in the derivative of logarithm of **partial likelihood function with respect to the given covariate***
(or: a difference between actual value of the given covariate for the given subject and expected value of particular covariate in a risk set).

ASSESSMENT OF MODEL ADEQUACY



Complex process of model assessment is divided into 5 steps:

ASSESSMENT OF MODEL ADEQUACY



Complex process of model assessment is divided into 5 steps:

1. Statistical Significance of Covariates

Likelihood Ratio Test, Score Test, Wald Test

ASSESSMENT OF MODEL ADEQUACY



Complex process of model assessment is divided into 5 steps:

1. Statistical Significance of Covariates

Likelihood Ratio Test, Score Test, Wald Test

2. Linear Relation between Covariates and Logarithm of Hazard

Plot of martingale residuals, Categorization of continuous variable

ASSESSMENT OF MODEL ADEQUACY



Complex process of model assessment is divided into 5 steps:

1. Statistical Significance of Covariates

Likelihood Ratio Test, Score Test, Wald Test

2. Linear Relation between Covariates and Logarithm of Hazard

Plot of martingale residuals, Categorization of continuous variable

3. Identification of Influential and Poorly Fitted Subjects

Plot of score residuals, dfbeta residuals, likelihood displacement statistics and $I - \max$ statistics

ASSESSMENT OF MODEL ADEQUACY



Complex process of model assessment is divided into 5 steps:

1. Statistical Significance of Covariates

Likelihood Ratio Test, Score Test, Wald Test

2. Linear Relation between Covariates and Logarithm of Hazard

Plot of martingale residuals, Categorization of continuous variable

3. Identification of Influential and Poorly Fitted Subjects

Plot of score residuals, dfbeta residuals, likelihood displacement statistics and $I - \max$ statistics

4. Proportional Hazard Assumption

Time – dependent variables, plot of Schoenfeld residuals

ASSESSMENT OF MODEL ADEQUACY



Complex process of model assessment is divided into 5 steps:

1. Statistical Significance of Covariates

Likelihood Ratio Test, Score Test, Wald Test

2. Linear Relation between Covariates and Logarithm of Hazard

Plot of martingale residuals, Categorization of continuous variable

3. Identification of Influential and Poorly Fitted Subjects

Plot of score residuals, dfbeta residuals, likelihood displacement statistics and $I - \max$ statistics

4. Proportional Hazard Assumption

Time – dependent variables, plot of Schoenfeld residuals

5. Overall Assessment of the Model Adequacy

Categorization of observation based on linear predictor value, plot of actual versus expected cumulative number of events

1. Statistical Significance of Covariates

- Partial likelihood ratio test

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

- Score test

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

- Wald test

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

ASSESSMENT OF MODEL ADEQUACY



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

Note: Censor = 0 indicates that event occurred (time variable contains full information), censor = 1 indicates that event did not occur (time variable is censored).

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

ASSESSMENT OF MODEL ADEQUACY



Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
AGE	1	-0.11147	0.04777	5.4442	0.0196	0.895
GENDER	1	1.87843	0.81161	5.3566	0.0206	6.543

*Both covariates are statistically significant,
both jointly and separately.*

2. Linear Relation between Covariates and Logarithm of Hazard

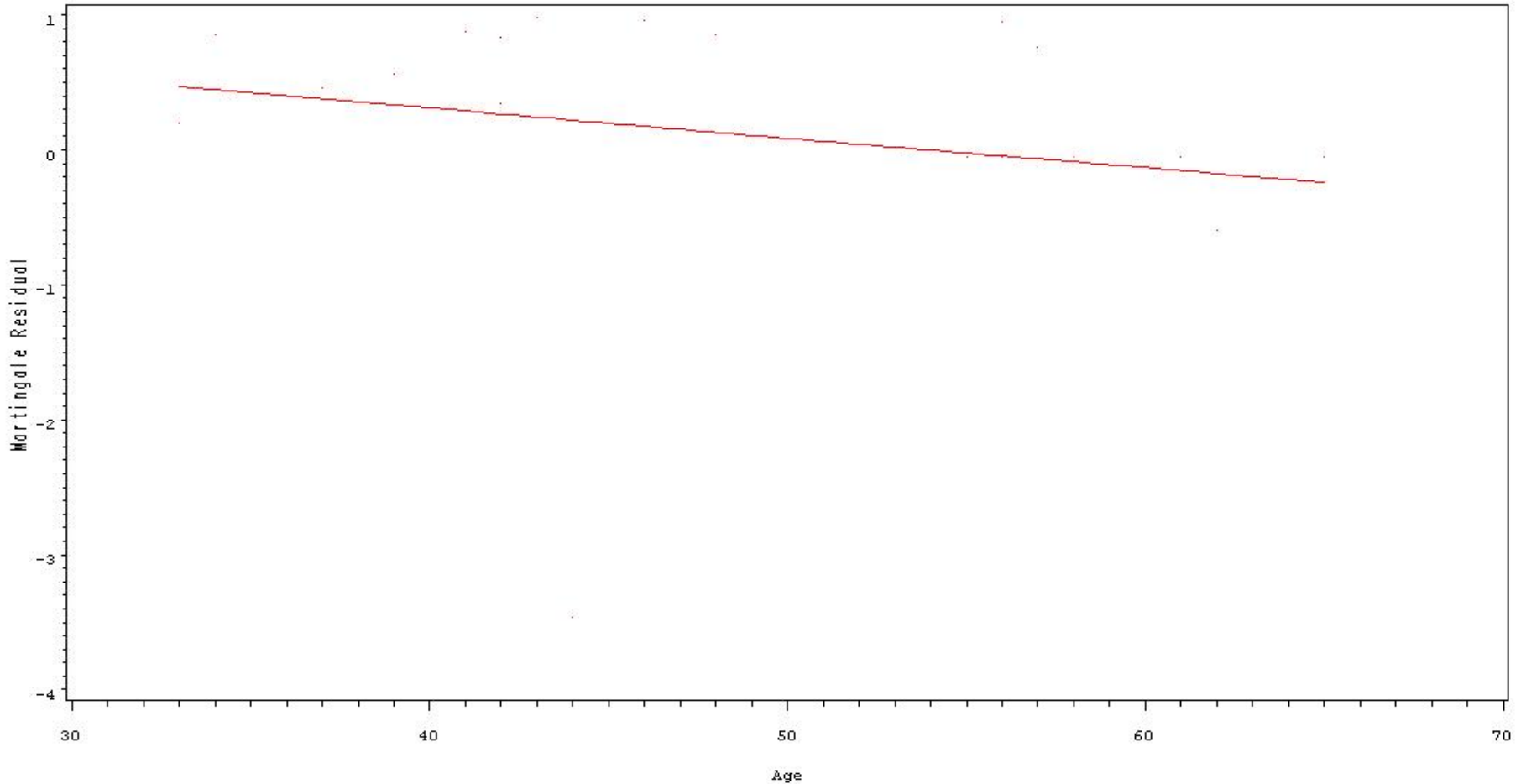
$$\lambda(t, \mathbf{z}) = \lambda_0(t) \cdot \exp(\beta \mathbf{z}) \quad \Rightarrow \quad \ln \lambda(t) = \ln \lambda_0(t) + \beta \mathbf{z}$$

➤ Plot of martingale residuals versus a covariate of interest

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

```
/* Plot of martingale residuals */  
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;
```

ASSESSMENT OF MODEL ADEQUACY



Line on the plot indicates type of relation between a covariate of interest (here: age) and logarithm of hazard; the above plot indicates linear relation.

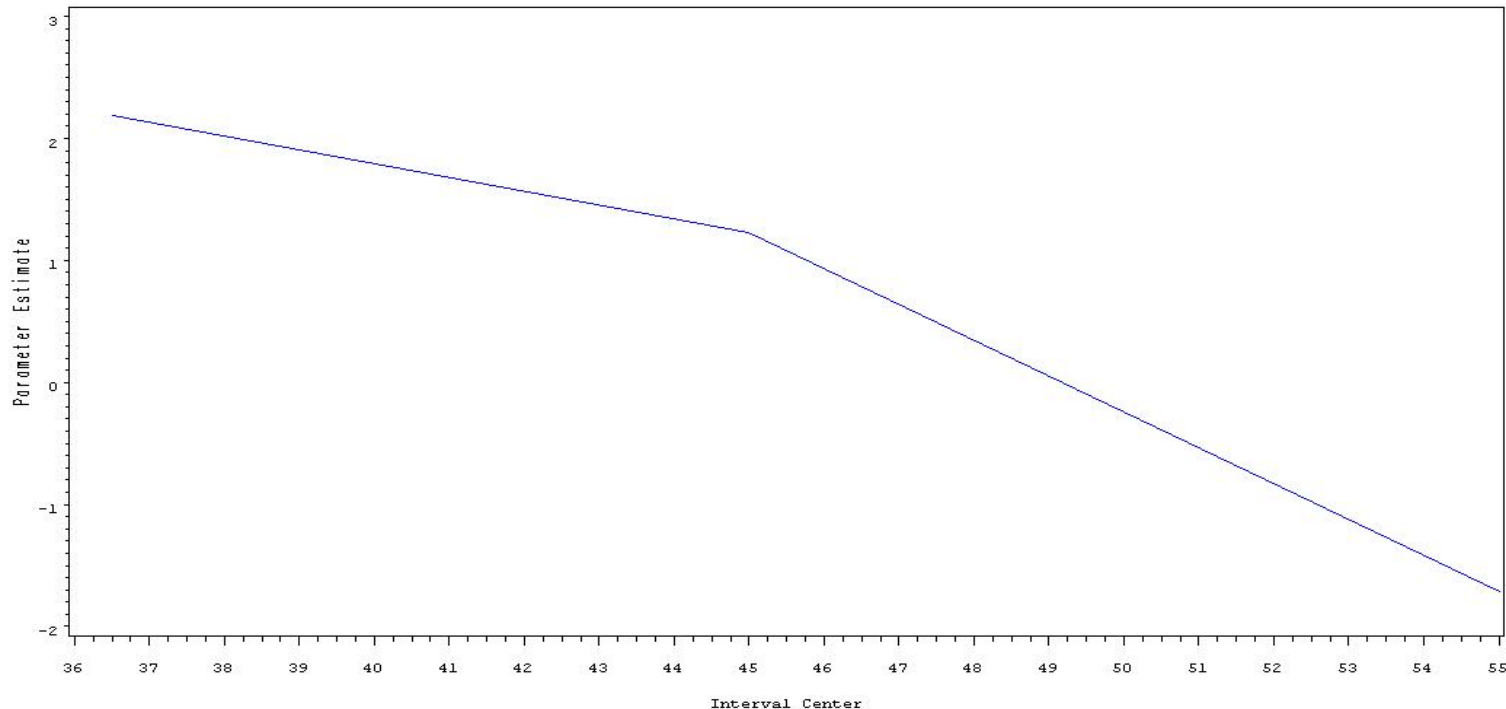
ASSESSMENT OF MODEL ADEQUACY



- Categorization of continuous variables and adding binary variables to the model – plot of parameters estimates versus centers of intervals

```
/* Model with additional binary variables */  
proc phreg data = sample outest = loglinear;  
  model time*censor(0) = gender w1 w2 w3 /ties =  
exact;  
run;
```

```
proc gplot data = loglinear;  
  ...  
run;
```



Plot of parameter estimates versus centers of intervals indicates type of relation between a covariate of interest (here: age) and logarithm of hazard; the above plot indicates linear relation.

ASSESSMENT OF MODEL ADEQUACY



3. Identification of Influential and Poorly Fitted Subjects

ASSESSMENT OF MODEL ADEQUACY



3. Identification of Influential and Poorly Fitted Subjects

- Plot of **score residuals** versus covariate of interest
*identification of subjects that **have value of the given covariate that differs from the sample average** to a great extent*

3. Identification of Influential and Poorly Fitted Subjects

- Plot of **score residuals** versus covariate of interest
*identification of subjects that **have value of the given covariate that differs from the sample average** to a great extent*
- Plot of **dfbeta residuals** versus covariate of interest
*identification of subjects that have strong **influence on parameters estimates***

3. Identification of Influential and Poorly Fitted Subjects

- Plot of **score residuals** versus covariate of interest
*identification of subjects that **have value of the given covariate that differs from the sample average** to a great extent*
- Plot of **dfbeta residuals** versus covariate of interest
*identification of subjects that have strong **influence on parameters estimates***
- Plot of **I – max statistics** and **likelihood displacement statistics** versus summary statistics, e.g. martingale residuals
*identification subjects that have strong **influence on the partial likelihood function value***

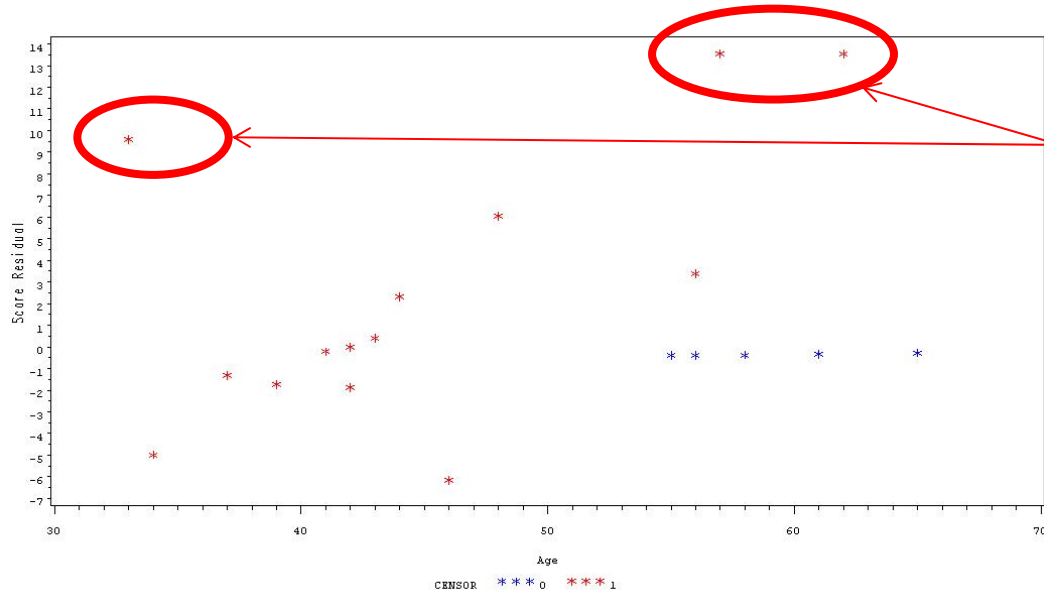
ASSESSMENT OF MODEL ADEQUACY



```
/* Model estimation with saving score, dfbeta and martingale
residuals as well as ld and likelihood displacement statistics */
proc phreg data = sample;
    model time*censor(0) = gender age / ties = exact;
    output out = score
    ressko = sc_gen sc_age /* Score residuals for each covariate */
    dfbeta = df_gen df_age /* Dfbeta residuals for each covariate */
    lmax = lmax /* L - max statistics */
    ld = ld /* Likelihood displacement statistics */
    resmart = resmart; /* Martingale residuals */
    id obs;
run;

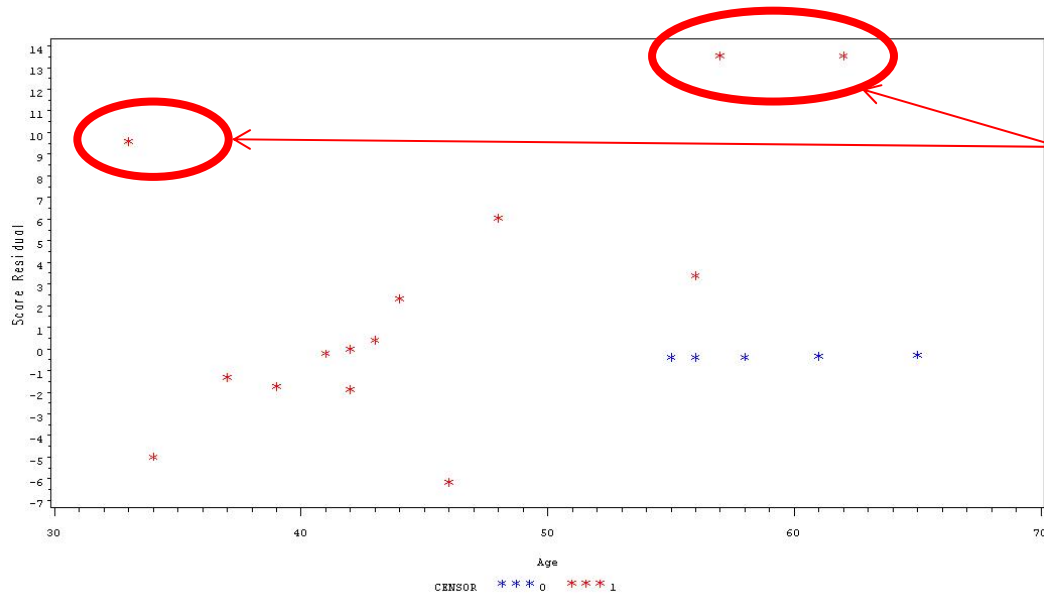
proc gplot data = score;
    ...
run;
```

ASSESSMENT OF MODEL ADEQUACY



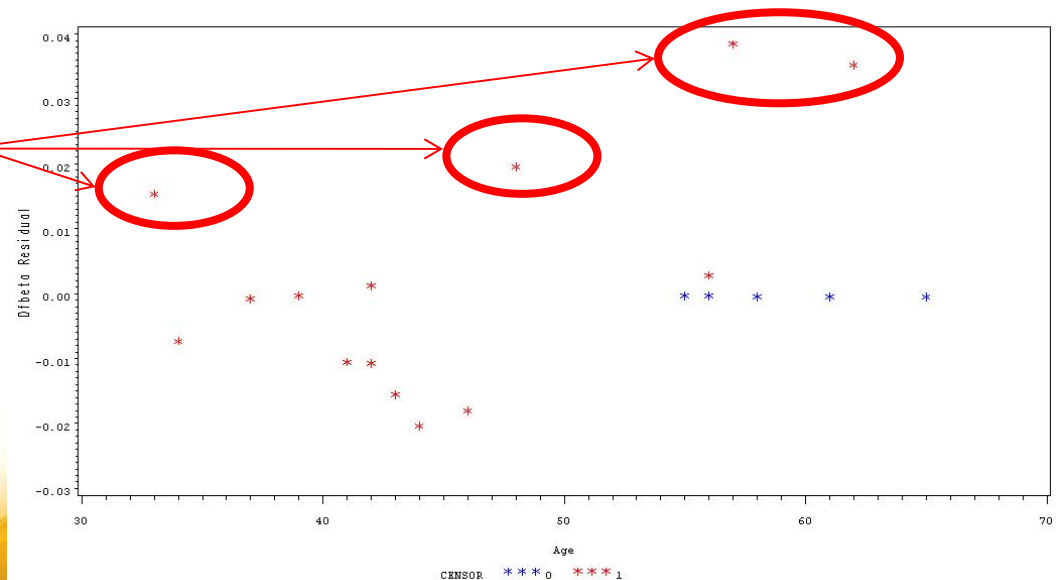
There are three subjects that seem to have value of variable age significantly higher than the sample average.

ASSESSMENT OF MODEL ADEQUACY

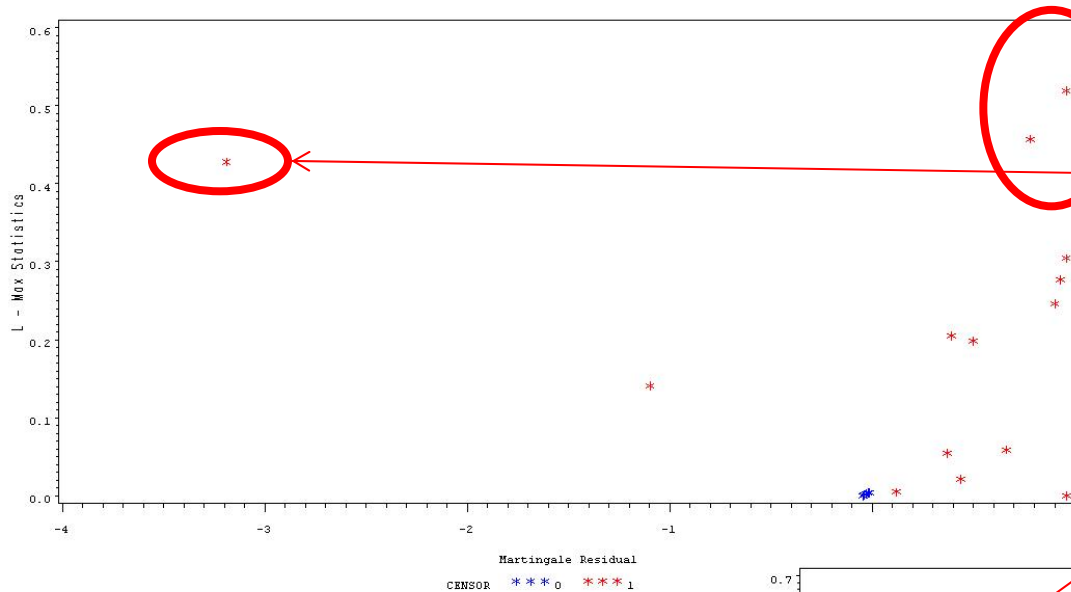


There are three subjects that seem to have value of variable age significantly higher than the sample average.

There are four subjects that seem to have strong influence on parameter estimate for variable age.

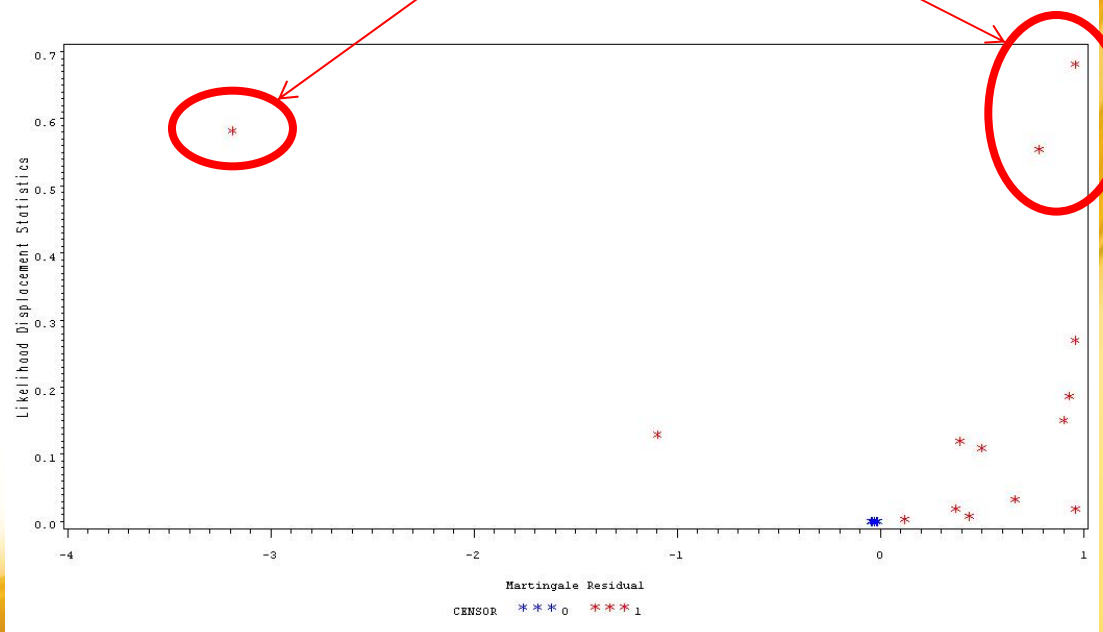


ASSESSMENT OF MODEL ADEQUACY



There are three subjects that seem to have strong influence on partial likelihood function.

Identified subjects need to be further investigated. The next step is reestimation of the model, excluding suspected observation and comparing new model with the original model. Finally, identified subjects may be excluded from the sample.



ASSESSMENT OF MODEL ADEQUACY



4. Proportional Hazard Assumption

➤ Time – dependent variables

```
/* Model estimation 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;
```

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
GENDER	1	10.76654	4.74708	5.1440	0.0233	47407.82
AGE	1	0.03012	0.19739	0.0233	0.8787	1.031
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

Time – dependent variables that were added to the model are not statistically significant which suggests that proportional hazard assumption is satisfied for both variables.

ASSESSMENT OF MODEL ADEQUACY

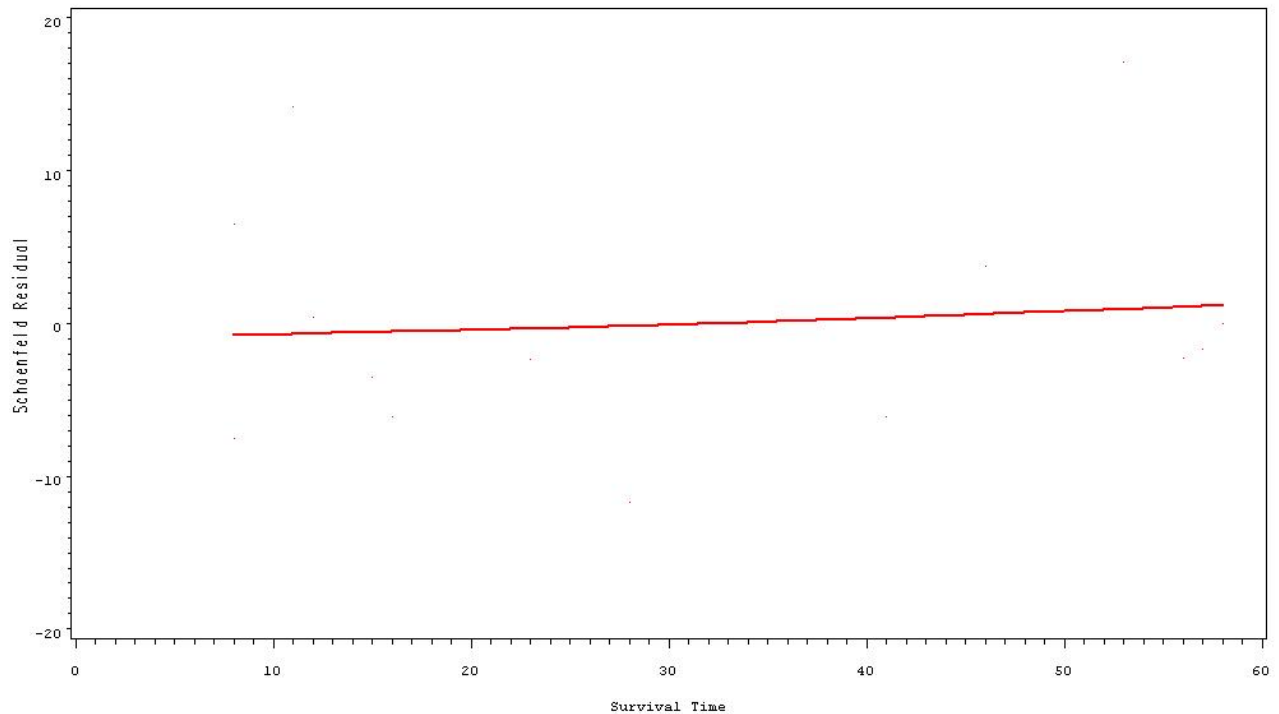


➤ Plot of Schoenfeld Residuals versus Time Variable

```
/* Model estimation with saving  
   Schoenfeld residuals */
```

```
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;
```



*Line on the plot
is approximately
horizontal which
suggests that
assumption
of proportional hazard
is satisfied.*

ASSESSMENT OF MODEL ADEQUACY



5. Overall Assessment of the Model Adequacy

➤ Categorization of observation based on linear predictor value

```
/* Linear predictor calculation */
data sample; set sample;
  xbeta = 1.87843 * gender - 0.11147 * age;
run;

/* Percentiles calculation */
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;
/* Binary variables*/
%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;

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

ASSESSMENT OF MODEL ADEQUACY



Analysis of Maximum Likelihood Estimates

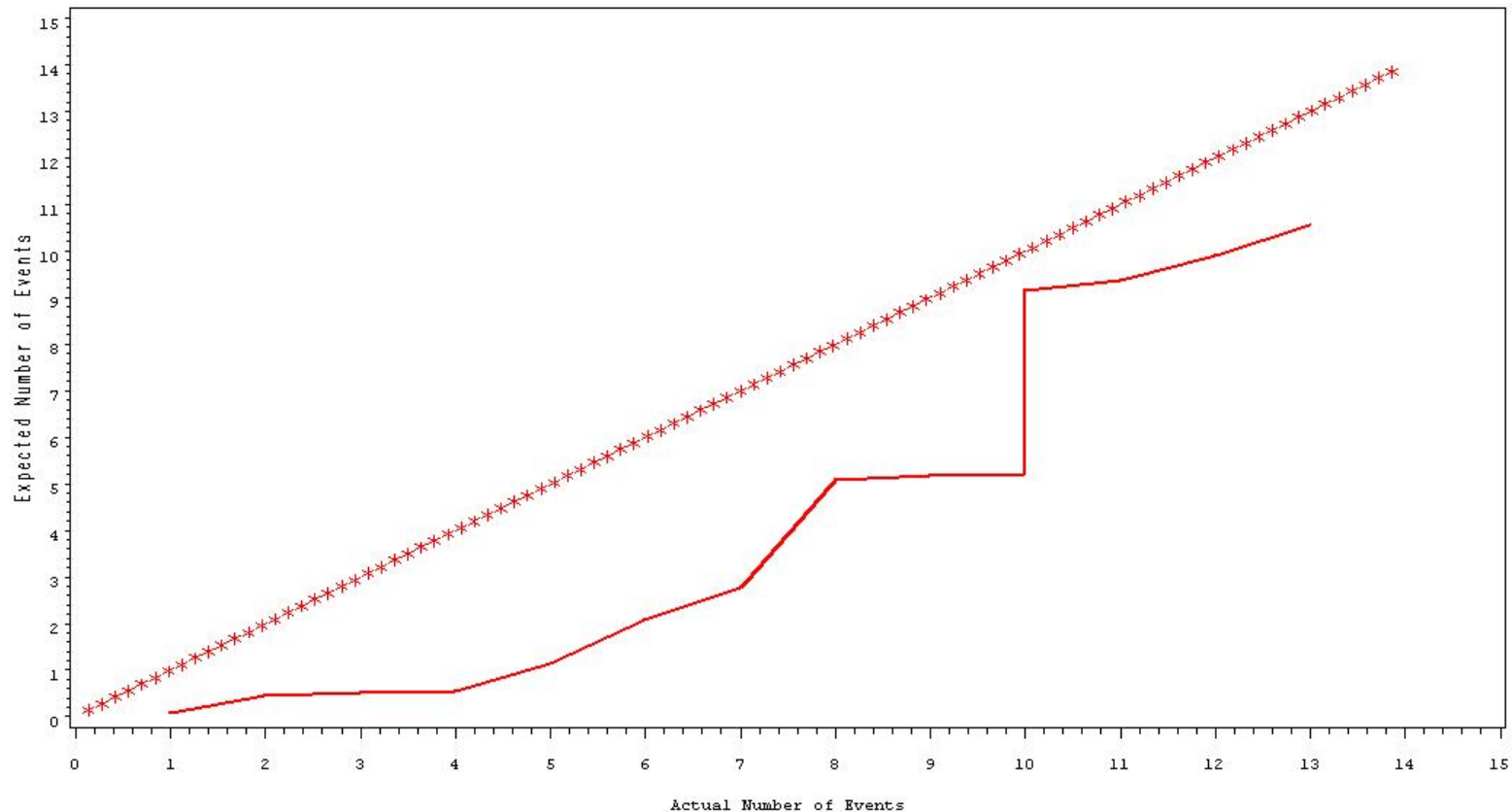
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
GENDER	1	0.79324	0.70054	1.2822	0.2575	2.211
AGE	1	-0.06102	0.06135	0.9893	0.3199	0.941
x10	1	-3.17656	1.65183	3.6981	0.0545	0.042
x20	0	0	.	.	.	.
x30	0	0	.	.	.	.
x40	0	0	.	.	.	.
x50	0	0	.	.	.	.
x60	0	0	.	.	.	.
x70	0	0	.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

None of the added binary variables is statistically significant at the level 0.05 which indicates well fit model.

ASSESSMENT OF MODEL ADEQUACY



- Plot of actual versus expected cumulative number of events



The line on the plot differs from a 45 degree line to a great extent which suggests that model specification should be reconsidered. It may be, among others, due to violations from model assumption for the other covariate: gender (as assumptions were examined only with respect to age) or outliers that were not excluded from the sample.

CONCLUSIONS



The aim of the current presentation was to underline the importance of process of model adequacy assessment which seems to be neglected sometimes.

It is crucial to follow the algorithm step by step and introduce necessary amendments in model specification if required.

All assumptions have to be satisfied with respect for all covariates and overall assessment of model positive before any statistical analyzes are performed on the basis of the model, otherwise they may result in misleading and improper conclusions.

Thanks for your attention!



CONTACT INFORMATION



Jadwiga Borucka

Quanticate Polska Sp. z o.o.

Hankiewiczza 2

02-103 Warsaw

Poland

Tel: +48(0) 22 576 21 40

Fax: +48(0) 22 576 21 59

E-mail: jadwiga.borucka@quanticate.com

Brand and product names are trademarks of their respective companies.