Statistický seminář

Coxův proporcionální hazardní model

Žaneta Miklová

Vysoká škola báňská - Technická univerzita Ostrava
Katedra aplikované matematiky

Ostrava, 21. ledna 2015

Analýza přežití

Analýza přežití

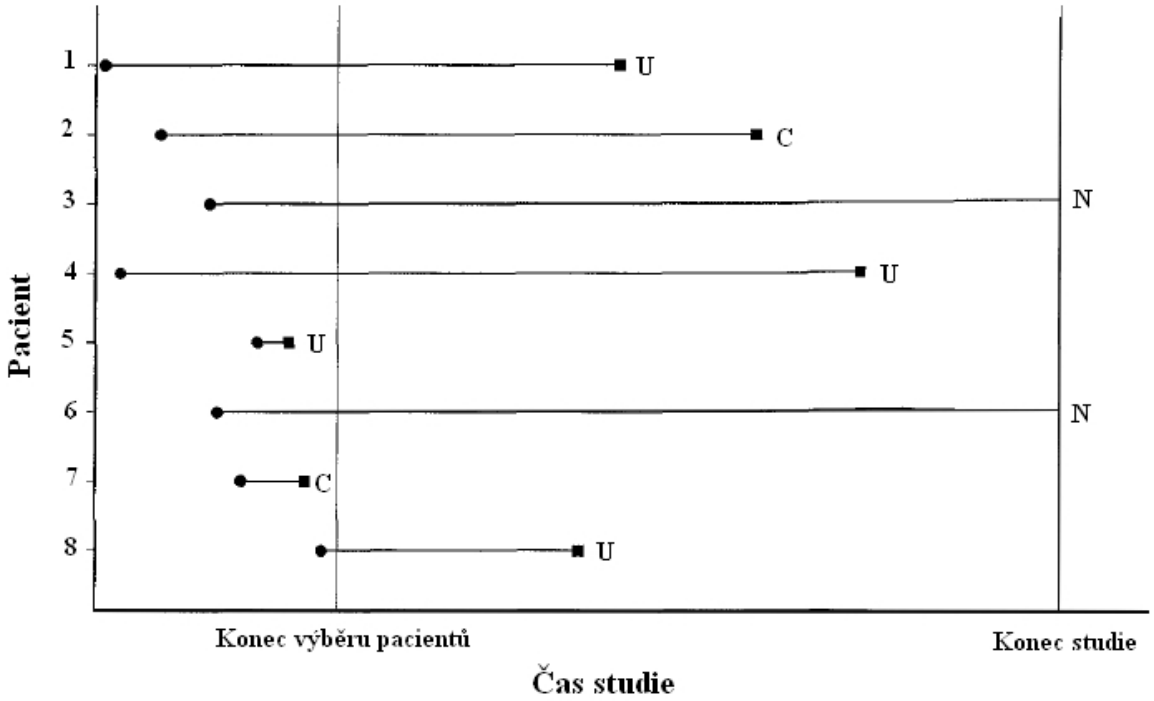
Analýza doby trvání do výskytu jedné nebo více událostí.

Terminologie:

- teorie spolehlivosti (mechanika, ekonomie)
- analýza historie (sociologie)
- analýza přežití (medicína)

Cenzorovaná data

- cenzorování časem (1. druhu)
- cenzorování výskytem událostí (2. druhu)
- náhodné cenzorování



Obrázek: Schéma času studie osmi pacientů

Funkce přežití

- vyjadřuje pravděpodobnost, že v intervalu $(0; t)$ nedojde k pooperačním komplikacím

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

Hazardní funkce

- nejedná se o pravděpodobnost
- ale o poměr hustoty pravděpodobnosti a funkce přežití

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}$$

- pravděpodobnost, že dojde k pooperačním komplikacím v následujícím krátkém úseku délky Δt za předpokladu, že k ní do času t nedošlo, je přibližně $h(t) \cdot \Delta t$

Kumulativní hazardní funkce

Častěji než s hazardní funkcí pracujeme s kumulativní hazardní funkcí.

$$H(t) = \int_0^t h(u)du$$

Věrohodnostní funkce

Nechť

- $X = (X_1, X_2, \ldots, X_n)$ je náhodný výběr
- $x = (x_1, x_2, \ldots, x_n)$ je jeho realizace
- $f(x, \Theta)$ je regulární hustota popisující populaci (Θ je neznámý parametr)

Potom věrohodnostní funkce je

$$\ell(\Theta|x) = \ell(\Theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i, \Theta)$$

- **rozdělení:** parametr je fixní, pozorování se mění
- **věrohodnostní funkce:** pozorování je fixní, parametr se mění

Modely v analýze přežití

1. třídění modelů

- parametrické
- neparametrické
- semiparametrické (*Coxův proporcionální model*)

2. třídění modelu

- AFT modely (*Accelerated Failure Time models*)
- PH modely (*Proportional Hazard models*)

AFT modely

Nechť u i-tého pacienta máme naměřené hodnoty $(t_i, c_i, x_1, \ldots, x_p)$, kde $i = 1, \ldots, n$. AFT modely definují model pro dobu do pooperačních komplikací T následovně

$$\begin{aligned} Y = ln(T) &= \beta_0 + \beta' x + \sigma\epsilon, \\ T &= e^{\beta_0 + \beta' x + \sigma\epsilon} = e^{\beta_0 + \sigma\epsilon} \cdot e^{\beta' x} \end{aligned}$$

Referenční doba do pooperačních komplikací T_0:

$$\begin{aligned} T_0 &= e^{\beta_0 + \sigma\epsilon} \\ T &= T_0 \cdot e^{\beta' x} \end{aligned}$$

Základní funkce přežití S_0:

$$S(t, x) = P(T > t|x) = P(T_0 e^{\beta' x} > t) = P(T_0 > te^{-\beta' x}) = S_0(te^{-\beta' x})$$

PH modely

PH modely respektují hazardní funkci, jejíž podobu dále ovlivňují vysvětlující proměnné.

Princip PH modelů:

$$h(t) \quad = \quad \Theta_0$$

PH modely

PH modely respektují hazardní funkci, jejíž podobu dále ovlivňují vysvětlující proměnné.

Princip PH modelů:

$$\begin{aligned} h(t) &= \Theta_0 \\ h(t) &= e^{\beta_0} \end{aligned}$$

PH modely

PH modely respektují hazardní funkci, jejíž podobu dále ovlivňují vysvětlující proměnné.

Princip PH modelů:

$$\begin{aligned} h(t) &= \Theta_0 \\ h(t) &= e^{\beta_0} \\ h(t) &= e^{\beta_0 + \beta' x} \end{aligned}$$

PH modely

PH modely respektují hazardní funkci, jejíž podobu dále ovlivňují vysvětlující proměnné.

Princip PH modelů:

$$\begin{aligned} h(t) &= \Theta_0 \\ h(t) &= e^{\beta_0} \\ h(t) &= e^{\beta_0 + \beta' x} \\ h(t, x, \beta) &= h_0(t) r(x, \beta) \end{aligned}$$

PH modely

PH modely respektují hazardní funkci, jejíž podobu dále ovlivňují vysvětlující proměnné.

Princip PH modelů:

$$\begin{aligned} h(t) &= \Theta_0 \\ h(t) &= e^{\beta_0} \\ h(t) &= e^{\beta_0 + \beta' x} \\ h(t, x, \beta) &= h_0(t) r(x, \beta) \end{aligned}$$

Poměr hazardních funkcí *(Hazard Ration)*:

$$\begin{aligned} HR(t, x_1, x_0) &= \frac{h(t, x_1, \beta)}{h(t, x_0, \beta)} \\ HR(t, x_1, x_0) &= \frac{h_0(t) r(x_1, \beta)}{h_0(t) r(x_0, \beta)} = \frac{r(x_1, \beta)}{r(x_0, \beta)} \end{aligned}$$

Exponencionální regresní model

Předpokládejme, že náhodná veličina T se řídí exponencionálním rozdělením $E(\lambda)$. Pak

$$S(t) = e^{-\lambda t}, \qquad h(t) = \lambda$$

Exponencionální regresní model

Předpokládejme, že náhodná veličina T se řídí exponencionálním rozdělením $E(\lambda)$. Pak

$$S(t) = e^{-\lambda t}, \qquad h(t) = \lambda$$

Neznámý parametr λ můžeme vyjádřit jako funkci p vysvětlujících proměnných $\lambda = e^{-(\beta_0+\beta_1 x+\cdots+\beta_p x_p)}$.

$$S(t) = e^{-te^{-(\beta_0+\beta_1 x)}}, \qquad S_0(t) = e^{-te^{-\beta_0}}$$

Exponencionální regresní model

Předpokládejme, že náhodná veličina T se řídí exponencionálním rozdělením $E(\lambda)$. Pak

$$S(t) = e^{-\lambda t}, \qquad h(t) = \lambda$$

Neznámý parametr λ můžeme vyjádřit jako funkci p vysvětlujících proměnných $\lambda = e^{-(\beta_0+\beta_1 x+\cdots+\beta_p x_p)}$.

$$S(t) = e^{-te^{-(\beta_0+\beta_1 x)}}, \qquad S_0(t) = e^{-te^{-\beta_0}}$$
$$h(t) = e^{-(\beta_0+\beta_1 x)}, \qquad h_0(t) = e^{-\beta_0}$$

Exponencionální regresní model

Předpokládejme, že náhodná veličina T se řídí exponencionálním rozdělením $E(\lambda)$. Pak

$$S(t) = e^{-\lambda t}, \qquad h(t) = \lambda$$

Neznámý parametr λ můžeme vyjádřit jako funkci p vysvětlujících proměnných $\lambda = e^{-(\beta_0+\beta_1 x+\cdots+\beta_p x_p)}$.

$$S(t) = e^{-te^{-(\beta_0+\beta_1 x)}}, \qquad S_0(t) = e^{-te^{-\beta_0}}$$
$$h(t) = e^{-(\beta_0+\beta_1 x)}, \qquad h_0(t) = e^{-\beta_0}$$

AFT tvar funkce přežití:

$$S(t) = e^{-te^{-(\beta_0+\beta_1 x)}} = e^{-(te^{-\beta_1 x})\cdot(e^{-\beta_0})} = S_0(te^{-\beta_1 x}).$$

PH tvar funkce přežití:

$$h(t) = e^{-(\beta_0+\beta_1 x)} = e^{-\beta_0}e^{-\beta_1 x} = h_0(t)e^{-\beta_1 x}$$

$$S(t) = e^{-te^{-(\beta_0+\beta_1 x)}} = e^{-te^{-\beta_0}e^{-\beta_1 x}} = \left(e^{-te^{-\beta_0}}\right)^{e^{-\beta_1 x}} = (S_0(t))^{e^{-\beta_1 x}}$$

Coxův proporcionální hazardní model

$$h(t) = h_0(t)e^{\beta' x}$$

Doba do pooperačních komplikací je spojitá náhodná veličina pocházející z rozdělení s hustotou pravděpodobnosti $f(t, x, \beta)$.

Obecně můžeme věrohodnostní funkci zapsat

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

Coxův proporcionální hazardní model

$$h(t) = h_0(t)e^{\beta' x}$$

Doba do pooperačních komplikací je spojitá náhodná veličina pocházející z rozdělení s hustotou pravděpodobnosti $f(t, x, \beta)$.

Obecně můžeme věrohodnostní funkci zapsat

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

$$L(\beta) = \sum_{i=1}^{n} \{ c_i ln [f(t_i, \beta, x_i)] + (1 - c_i) ln [S(t_i, \beta, x_i)] \}$$

Coxův proporcionální hazardní model

$$h(t) = h_0(t)e^{\beta' x}$$

Doba do pooperačních komplikací je spojitá náhodná veličina pocházející z rozdělení s hustotou pravděpodobnosti $f(t, x, \beta)$.

Obecně můžeme věrohodnostní funkci zapsat

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

$$L(\beta) = \sum_{i=1}^{n} \{ c_i ln [f(t_i, \beta, x_i)] + (1 - c_i) ln [S(t_i, \beta, x_i)] \}$$

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

Částečná věrohodnostní funkce

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

Odhad rozptylu a **standardní chyby** odhadu koeficientů:

$$\hat{D}\hat{\beta} = I(\hat{\beta})^{-1}, \qquad \hat{SE}(\hat{\beta}) = \sqrt{\hat{D}\hat{\beta}}$$

$$I(\beta) = -\frac{\partial^2 L_p(\beta)}{\partial \beta^2}$$

Opakující se pozorované časy v datech:

- přesná metoda
- Breslowova aproximace
- Efronova aproximace
- Coxova aproximace

Odhady parametru β

- MLE
- Newton-Rhapsonova metoda (iterační metoda)

Odhady parametru β

- MLE
- Newton-Rhapsonova metoda (iterační metoda)

Významnost parametrů β:

- intervaly spolehlivosti pro β
- Waldova statistika
- Skóre test
- test poměru částečné věrohodnostní funkce

Test poměru částečné věrohodnostní funkce

$$G = 2\left\{L(\hat{\beta}) - L(0)\right\}$$

Statistika G se řídí χ^2 rozdělením s p stupni volnosti (za každý odhadovaný parametr jeden stupeň volnosti). Testujeme nulovou hypotézu

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$
H_A : Aspoň jeden z koeficientů je různý od nuly.

$$p - hodnota = P\left(\chi^2(p) \geq 2\left\{L(\hat{\beta}) - L(0)\right\}\right)$$

Popis dat

- doba do pooperačních komplikací (*uvedeno ve dnech*)
- cenzorovaný údaj (*0-cenzorovaný, 1-necenzorovaný*)
- skupina (*odpovídá druhu operace, 0-otevřená, 1-laparoskopická*)
- věk
- pohlaví (*0-muž, 1-žena*)
- diagnóza (*0-c18, 1-c19, 2-c20*)
- BMI (*Body Mass Index*)
- ASA (*American Society of Anaesthesiology Classification, kódováno hodnotami 1-4*) aj.

	n	V procentech
Necenzorované časy $c = 1$	426	54.3%
Cenzorované časy $c = 0$	359	45.7%
Celkem	785	100.0%

Číselná vs. Kategoriální proměnná:

úroveň		ASA_1	ASA_2	ASA_3
ASA	1	0	0	0
	2	1	0	0
	3	0	1	0
	4	0	0	1

Číselná vs. Kategoriální proměnná:

úroveň		ASA_1	ASA_2	ASA_3
ASA	1	0	0	0
	2	1	0	0
	3	0	1	0
	4	0	0	1

Sestavení modelu:

Základem je statistika založená na **testu poměru částečné věrohodnostní funkce**, která se označuje jako $-2ln(\ell(\hat{\beta}))$.

- dopředný výběr
- zpětná eliminace
- obecná strategie

Model	$-2ln\ell(\hat{\beta})$	**Pokles**	**df**	**p-hodnota**
nulový	5279.443			
stadium	5005.904	273.539	3	0
ASA	5256.695	22.748	3	< 0.001
věk	5262.062	17.381	1	< 0.001
ICHS	5262.593	16.850	1	< 0.001
DM	5268.932	10.511	1	0.001
grading	5266.918	12.525	2	0.002
BMI	5270.970	8.473	1	0.004
délka operace	5272.351	7.092	1	0.008
arytmie	5272.443	7.000	1	0.008
hypertenze	5275.212	4.231	1	0.040
cebrovaskulární	5275.896	3.547	1	0.060
krevní ztráta	5276.217	3.226	1	0.072
perop. komplikace	5277.006	2.437	1	0.119
pulmonální	5278.181	1.262	1	0.261
...				

Model	$-2ln\ell(\hat{\beta})$	**Nárůst**	**df**	**p-hodnota**
S+AS+V+I+G+D+B+DM+A+H	4927.636			
-*stadium* (*S*)	5217.911	290.275	3	0
-*grading* (*G*)	4938.122	10.486	2	0.005
-*věk* (*V*)	4934.398	6.762	1	0.009
-*arytmie* (*A*)	4934.188	6.552	1	0.010
-*BMI* (*B*)	4931.449	3.813	1	0.051
-*délka operace* (*D*)	4930.506	2.870	1	0.090
-*ASA* (*AS*)	4933.654	6.018	3	0.111
-*DM*	4929.679	2.043	1	0.153
-*hypertenze* (*H*)	4928.223	0.587	1	0.444
-*ICHS* (*I*)	4927.846	0.210	1	0.647

Model	$-2ln\ell(\hat{\beta})$	**Nárůst**	**df**	**p-hodnota**
stadium+grading+věk+arytmie	4947.392			
-stadium	5245.556	298.164	3	0
-věk	4972.797	25.405	1	< 0.001
-arytmie	4958.492	11.100	1	0.001
-grading	4957.265	9.873	2	0.007

Model	$-2ln\ell(\hat{\beta})$	**Pokles**	**df**	**p-hodnota**
stadium+věk+grading+arytmie	4947.392			
krevní ztráta	4943.805	3.587	1	0.058
cebrovaskulární	4943.798	3.594	1	0.058
pulmonální	4945.087	2.305	1	0.129
perop. komplikace	4945.610	1.782	1	0.182
diagnóza	4944.176	3.216	2	0.200
jaterní	4945.854	1.538	1	0.215
konverze	4944.638	2.754	2	0.252
renální	4946.844	0.548	1	0.459
předchozí operace	4947.191	0.201	1	0.654
pohlaví	4947.304	0.088	1	0.767

Model	$-2ln\ell(\hat{\beta})$	**Pokles**	**df**	**p-hodnota**
stadium+věk+grading+arytmie	4947.392			
skupina	4945.889	1.503	1	0.220

To, které proměnné budou zahrnuty do modelu, závisí značně na volbě strategie. Např. pokud bychom zvolili metodu **dopředního výběru**, pak by výsledný model zahrnoval proměnné *stadium*, *věk*, *arytmie*, *grading*, *ASA*, *krevní ztráta* a *délka operace*.

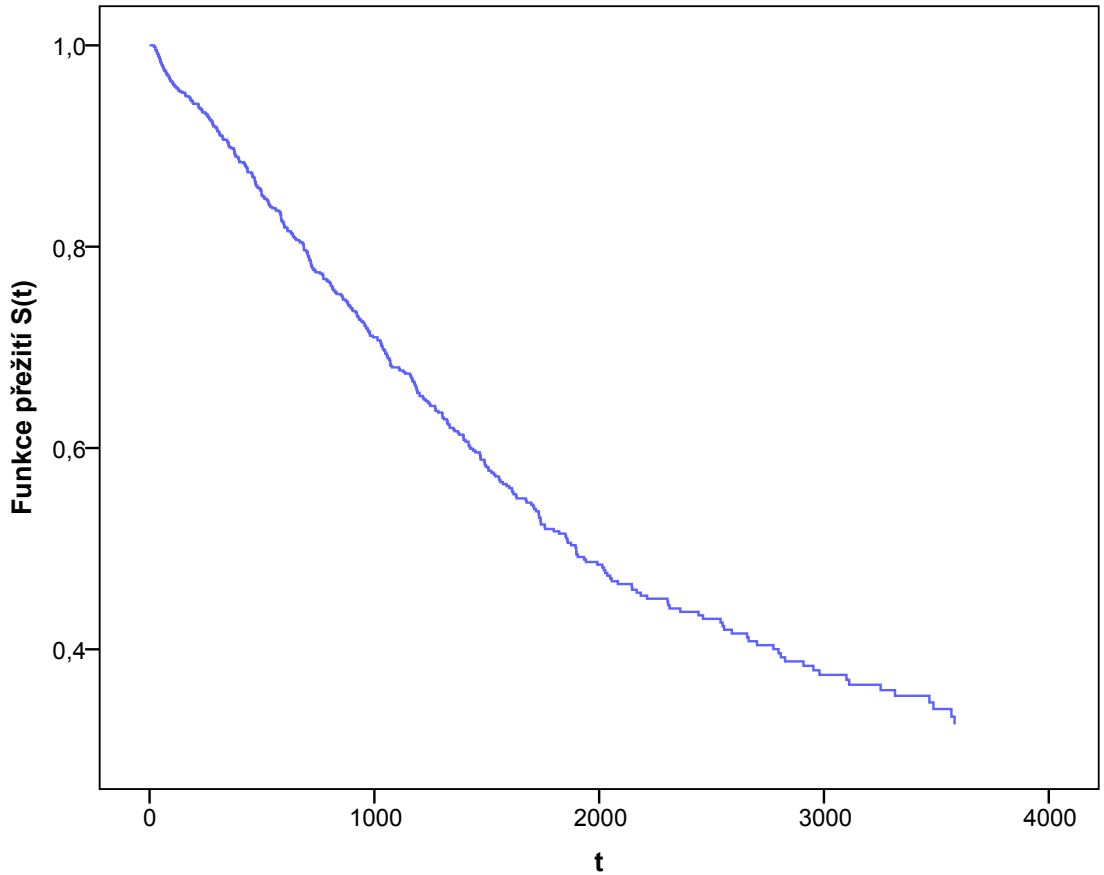
Závěry vyplývající z modelu

	$\hat{\beta}$	$SE(\hat{\beta})$	$e^{\hat{\beta}}$	95% CI pro $e^{\hat{\beta}}$ dolní	horní
$stadium_1$	0.472	0.208	1.603	1.066	2.410
$stadium_2$	1.180	0.198	3.254	2.207	4.798
$stadium_3$	2.560	0.199	12.936	8.754	19.106
věk	0.026	0.005	1.026	1.016	1.036
arytmie	0.517	0.147	1.676	1.256	2.237
$grading_1$	-0.183	0.109	0.833	0.673	1.031
$grading_2$	0.278	0.159	1.321	0.969	1.801

Výsledný model má tvar:

$$h(t) = h_0(t)e^{0.472ST_1+1.180ST_2+2.560ST_3+0.517AR+0.026vek-0.183GR_1+0.278GR_2},$$

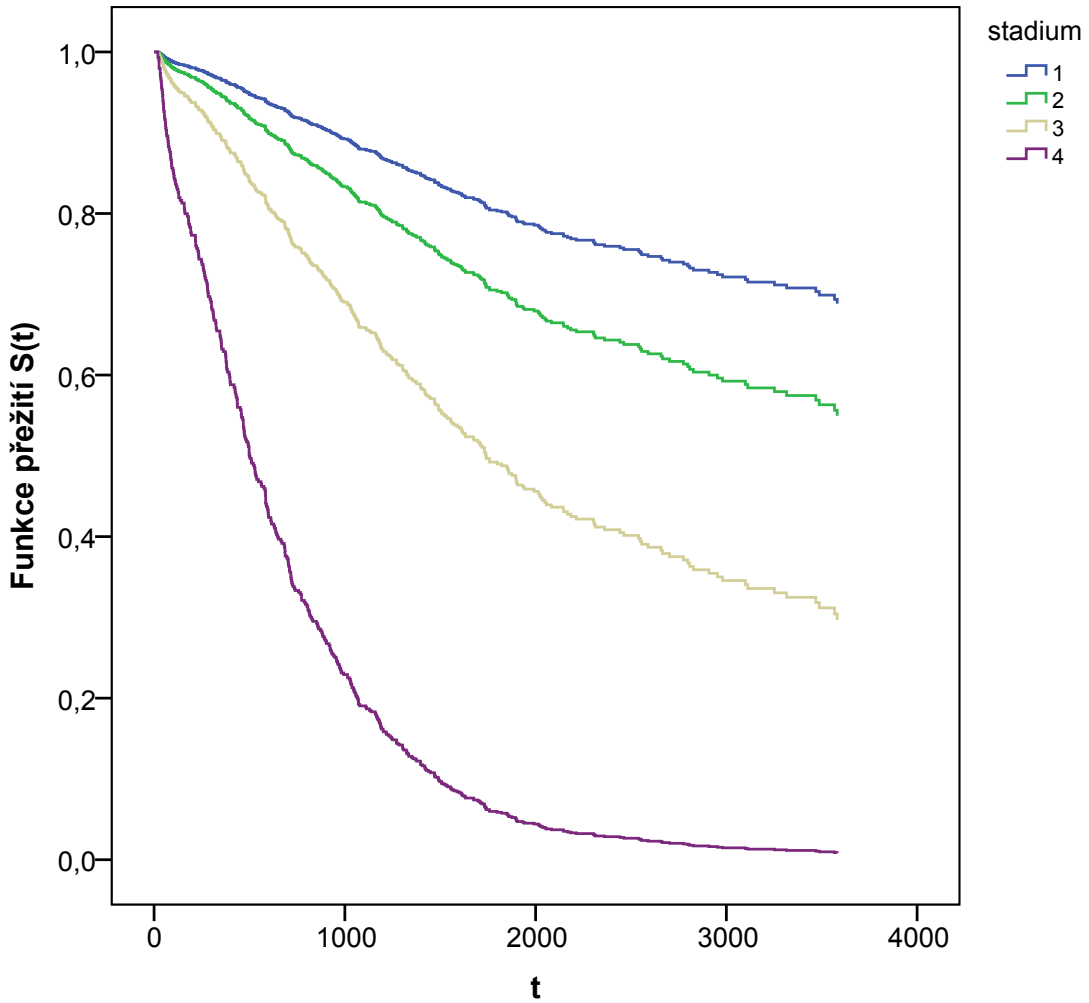
	$stadium_1$	$stadium_2$	$stadium_3$	*věk*	*arytmie*	$grading_1$	$grading_2$
Průměr	0.284	0.311	0.234	65.403	0.116	0.582	0.102



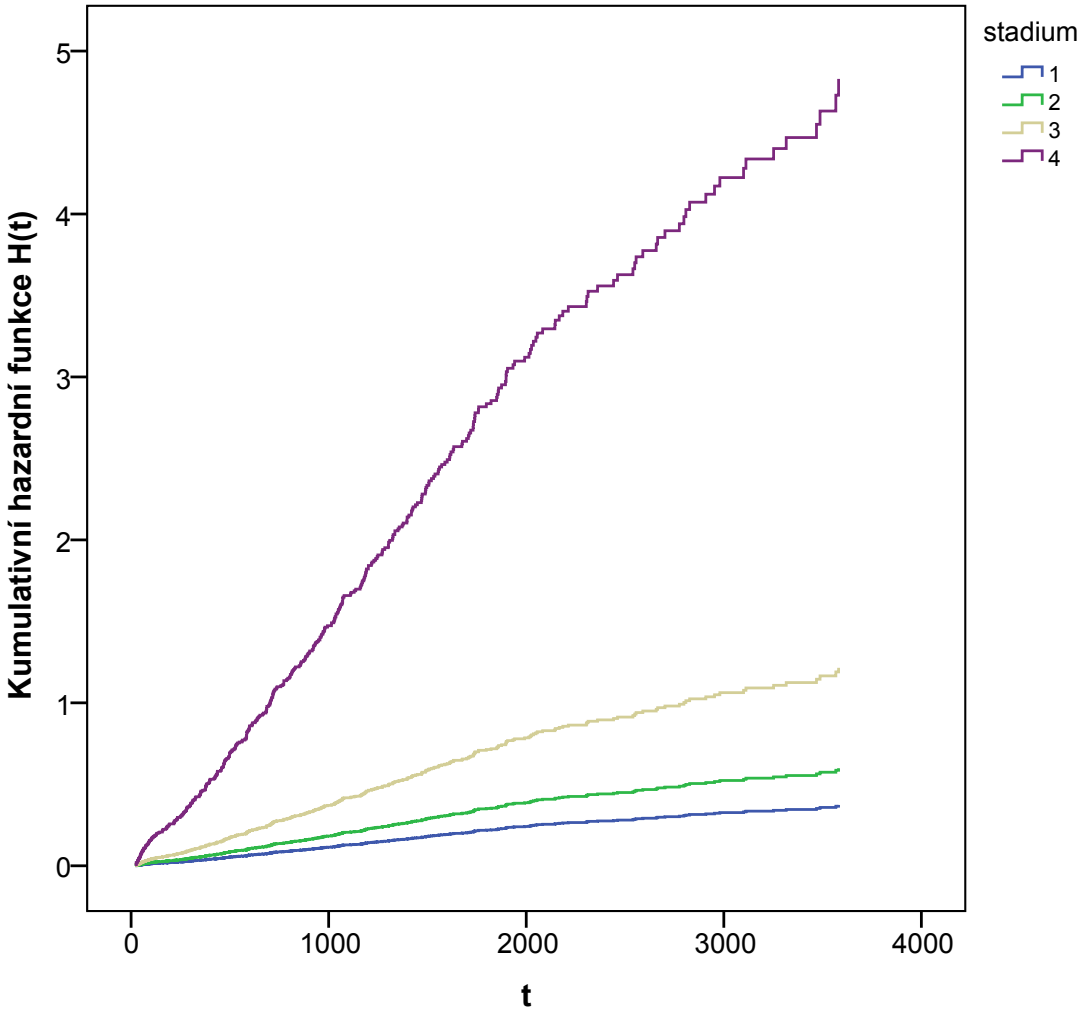
Vliv jednotlivých vysvětlujících proměnných - *stadium*

		Vzor			
	Průměr	1	2	3	4
$stadium_1$	0.284	0.000	1.000	0.000	0.000
$stadium_2$	0.311	0.000	0.000	1.000	0.000
$stadium_3$	0.234	0.000	0.000	0.000	1.000
věk	65.403	65.403	65.403	65.403	65.403
arytmie	0.116	0.116	0.116	0.116	0.116
$grading_1$	0.582	0.582	0.582	0.582	0.582
$grading_2$	0.102	0.102	0.102	0.102	0.102

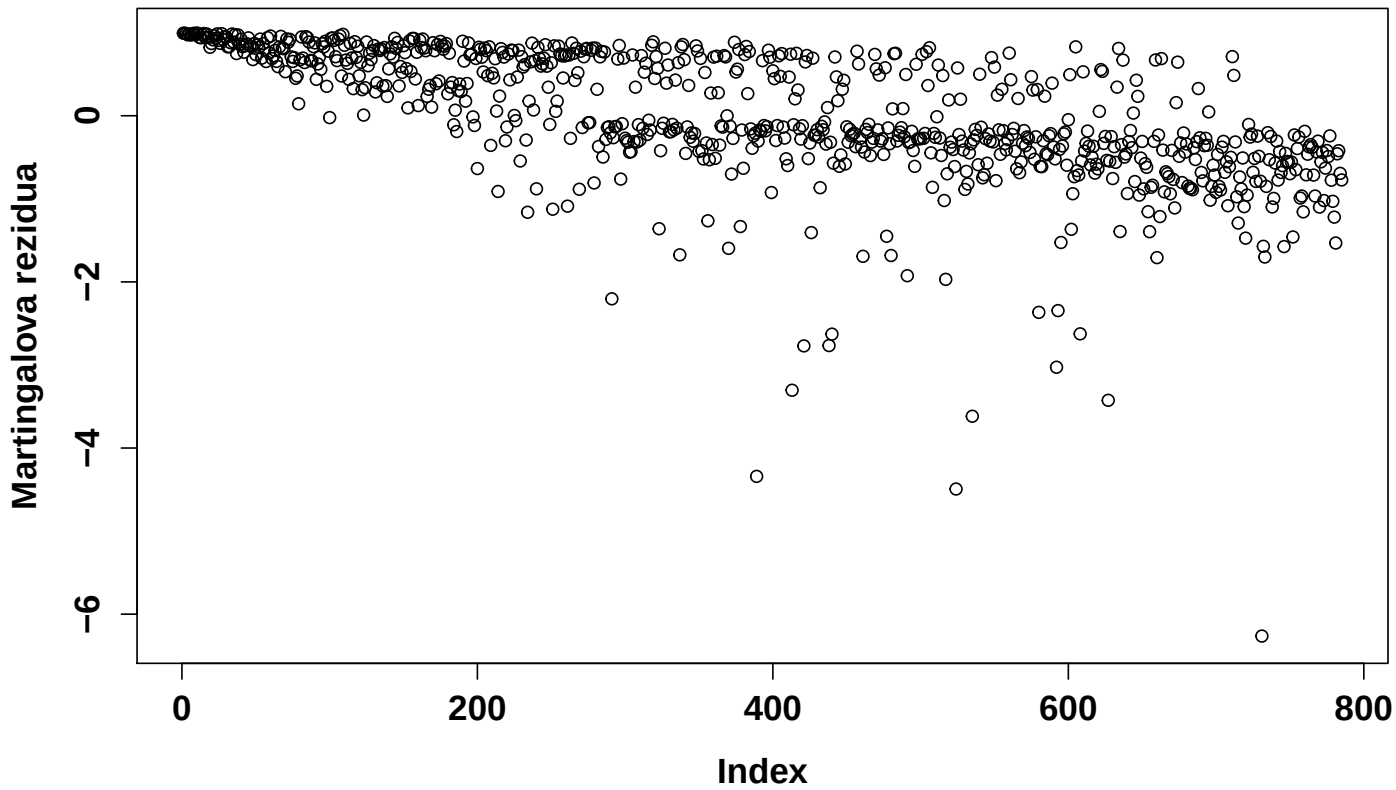
Vliv jednotlivých vysvětlujících proměnných - *stadium*



Vliv jednotlivých vysvětlujících proměnných - *stadium*

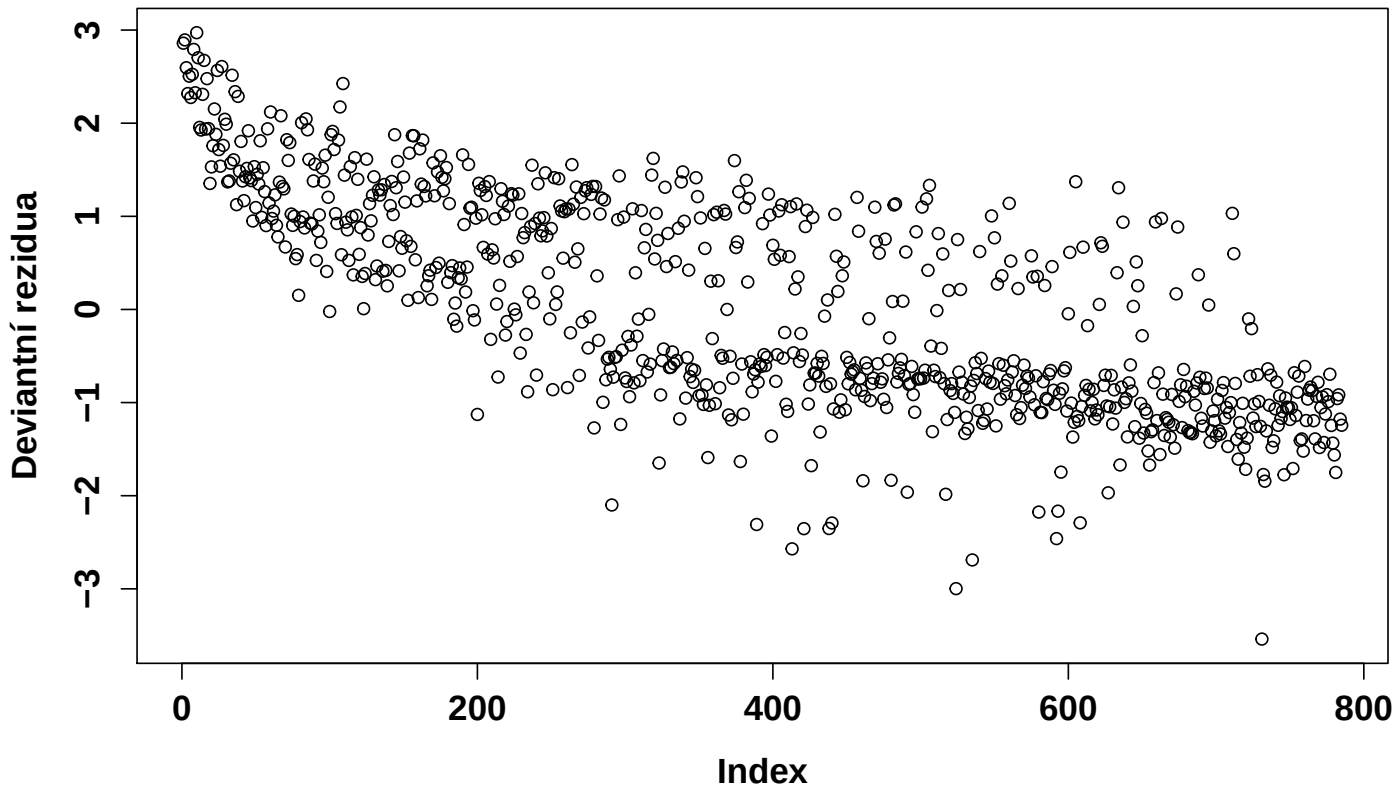


Ověření adekvátnosti modelu - Martingalova rezidua



Obrázek: Martingalova rezidua

Ověření adekvátnosti modelu - Deviantní rezidua



Obrázek: Deviantní rezidua

Závěr

- nejvýznamnější vysvětlující proměnné *stadium*, *arytmie*, *grading* a *věk*
- nejvíce hazardní funkci ovlivňuje proměnná *stadium* (riziko pooperačních komplikací se může navýšit třináctkrát)
- proměnná *skupina* nemá statisticky významný vliv na změnu hazardní funkce (laparoskopická vs. otevřená)

Děkuji za pozornost.