UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

Diplomová práce

Coxův model proporcionálních rizik a jeho diagnostika

Vedoucí bakalářské práce: **Ondřej Vencálek, Ph.D.**
Vypracoval: **Michal Polák**
Studijní program: B1103 Aplikovaná matematika
Studijní obor Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2016

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Michal Polák

Název práce: Coxův model proporcionálních rizik a jeho diagnostika

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: Mgr. Ondřej Vencálek, Ph.D.

Rok obhajoby práce: 2016

Abstrakt: Tato práce popisuje základní teorii Coxova modelu proporcionálních rizik a jeho diagnostiku. Cílem práce je nastudovat a aplikovat některé diagnostické nástroje pro ověření určitých předpokladů Coxova modelu. V práci se věnujeme především ověření proporcionality hazardu pomocí Schoenfeldových reziduí, ale jsou zde popsaný i další diagnostické nástroje Coxova modelu.

Klíčová slova: Analýza přežívání, Coxův model, Schoenfeldova rezidua, Martingalová rezidua, hazardní funkce, předpoklad proporcionality

Počet stran: 56

Počet příloh: 0

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Michal Polák

Title: Cox proportional hazard model and it‘s diagnostics

Type of thesis: Master's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: Mgr. Ondřej Vencálek, Ph.D.

The year of presentation: 2016

Abstract: This thesis describes basic theory of Cox proportional hazard model and it‘s diagnostics. The goal is to study and apply some diagnostic tools for verification of certain assumptions of Cox proportional hazard model. In the thesis we primarily concentrate on checking of proportional hazard assumption using Schoenfeld residuals, moreover we describe other diagnostic tools for Cox proportional hazard model.

Key words: Survival analysis, Cox proportional hazard model, Schoenfeld residuals, Martingale residuals, hazard function, assumption of proportionality

Number of pages: 56

Number of appendices: 0

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracoval samostatně pod vedením pana Mgr. Ondřeje Vencálka, PhD. s použitím uvedené literatury.

V Olomouci dne

Obsah

Poděkování

Rád bych poděkoval zejména vedoucímu mé diplomové práce panu Mgr. Ondřeji Vencálkovi PhD. za cenné rady, spolupráci a všechen čas, který mi věnoval během konzultací. Velký dík patří také mé rodině, která mě po celou dobu studia podporovala.

Úvod

Cílem této práce je srozumitelně popsat teorii Coxova modelu proporcionálních rizik a jeho diagnostických nástrojů. A to způsobem, aby čtenář bez znalosti teorie martingalů byl schopen textu porozumět.

V samotném úvodu představíme datovou sadou, na kterou budeme průběžně aplikovat vybrané metody.

V první kapitole se budeme zabývat klíčovými pojmy a základní teorií analýzy přežívání potřebné ke konstrukci Coxova modelu. V další části vysvětlíme zavedení, předpoklady a interpretaci Coxova modelu proporcionálních rizik. Detailně popíšeme konstrukci parciální věrohodnostní funkce a ukážeme si, proč není možné použít v regresi běžně používanou věrohodnostní funkci. V této sekci zmíníme i některé testy významnosti regresních parametrů.

V další kapitole se zaměřím na diagnostiku Coxova modelu pomocí různých typů reziduí. Především se zaměříme na diagnostiku klíčového předpokladu proporcionality hazardu. K tomuto účelu představíme Schoenfeldova rezidua a normovaná Schoenfeldova rezidua, pomocí nichž tento předpoklad ověřujeme. K ověření proporcionality použijeme Grambschové-Therneauv test a také jeho grafickou verzi. Dále si ukážeme, jak lze pomocí martingalových reziduí určit typ funkční závislosti hazardní funkce na spojitém regresoru. Pro tuto diagnostickou metodu jsme zkonstruovali simulaci založenou na generování hodnot z binomického rozdělení. Pomocí simulace prověříme diagnostické vlastnosti této metody. Na závěr si ukážeme, jakým způsobem se provádí detekce odlehlých pozorování pomocí deviančních reziduí.

Veškeré výpočty a grafy uvedené v této práci budou vytvořeny prostřednictvím statistického softwaru R.

1. Datová sada

Ve své práci budu vybrané metody aplikovat na reálnou datovou sadu, jejíž původ z důvodu citlivosti dat nebude uveden. Cílem je zkoumat vliv daných regresorů na dobu přežití pacienta. Událostí, kterou v tomto případě pozorujeme je úmrtí pacienta v důsledku rakovinového onemocnění. Počateční okamžik, kdy pacient vstupuje do studie je operace pacienta.

Data										
ID	survival	event	Věk	Transfůze	Chemoterapie	Stádium	Marker1	Marker2	Marker3	Marker4
1	36.34	0	66	0	0	1	0	0	0	0
2	38.47	0	62	0	1	1	1	0	1	0
3	7.62	1	54	1	1	2	1	0	0	0
4	41.92	0	77	0	0	1	0	0	0	0
5	41.52	0	77	0	0	1	0	0	0	0
6	42.42	0	70	0	0	1	0	1	0	0
7	45.47	0	68	0	1	2	0	0	0	0
8	40.84	0	60	1	0	1	0	0	0	0

Tabulka 1: Ukázka datové sady

Proměnné

- ID - identifikační číslo pacienta
- survival - spojitá veličina označující dobu přežití v týdnech
- event - proměnná rozlišující zda došlo u pacienta k selhání (event=1) nebo cenzorování (event=0)
- Věk - věk pacienta v době operace
- Transfúze - logická proměnná označující, zda pacientovi byla podána transfúze
- Chemoterapie - logická proměnná označující, zda pacient absolvoval chemoterapii
- Stádium - ordinální proměnná označující velikost nádoru, která má celkem 5 kategorií
- Marker1 - logická proměnná označující přítomnost určitého markeru v krvi
- Marker2 - logická proměnná označující přítomnost určitého markeru v krvi

- Marker3 - logická proměnná označující přítomnost určitého markeru v krvi

- Marker4 - logická proměnná označující přítomnost určitého markeru v krvi

2. Coxův model proporcionálních rizik

V této kapitole uvedu základní pojmy analýzy přežití, které budou později nezbytné pro konstrukci Coxova modelu. Definice doplním o slovní komentář, který by měl čtenáři pomoci s porozuměním a interpretací. Dále zavedu model proporcionálních rizik a zaměřím se na jeho předpoklady. Podíváme se také na odhad modelu z datové sady a interpretaci výsledků. V závěru se pokusím přiblížit, jak v tomto modelu probíhá inference, která se od klasické regrese liší. V této kapitole bylo čerpáno z [**?**], [**?**],[**?**] a [**?**].

2.1. Základní teorie analýzy přežívání

Klíčovým pojmem pro analýzu přežití je nezáporná náhodná veličina T, která bývá v literatuře označována jako *doba přežití*. Tato veličina uvádí dobu od počátku do okamžiku selhání v dané jednotce času. Co rozumíme počátkem a selháním musí být definováno samotnou studií, pro kterou analýzu přežívání využíváme.

Dalším důležitým pojmem je princip cenzorování. Ve své práci budu zohledňovat pouze jeden typ cenzorování a to princip cenzorování

zprava. Při analýze přežívání se v naší studii mohou vyskytovat jedinci, u nichž nedošlo k selhání během období, kdy ve studii působili. Tato data podléhají tzv. mechanismu cenzorování zprava. Uvažujme studii, jejíž počátek je definován jako první den v měsíci a která končí posledním dnem tohoto měsíce. Prvním příkladem cenzorování zprava může být jedinec, u něhož během tohoto období nedošlo k selhání. Druhým příkladem může být jedinec, který ze studie vypadne např. patnáctý den v měsíci z jakékoliv jiné příčiny než je definice selhání. Oba tyto subjekty jsou neselhavší a jsou tedy zprava cenzorovány.

K tomu, aby byly metody analýzy přežití platné, je třeba, aby mechanismus cenzorování byl nezávislý.

Pravé cenzorovací schéma je *nezávislé*, jestliže je riziko každého (neselhavšího a necenzorovaného) jedince v každém čase $t > 0$ stejné, jako by bylo, kdyby žádné cenzorování nenastalo. Nechť $Y(t) = 1$ znamená, že jedinec do času t nebyl cenzorován. Matematický zápis této nezávislosti je

$$\lim_{h\to 0+} \frac{P(T \in [t, t+h] | T \geq t, \boldsymbol{x})}{h} = \lim_{h\to 0+} \frac{P(T \in [t, t+h] | T \geq t, \boldsymbol{x}, Y(t) = 1)}{h}.$$

Obecně lze říct, že cenzorovací schéma je nezávislé, jestliže pravděpodobnost cenzorování daného jedince v okamžiku t nezávisí na pravděpodobnosti selhání v tomto čase, naopak mechanizmus cenzorování může záviset na vysvětlujících proměnných $\boldsymbol{x}$ nebo náhodných procesech nezávislých na časech selhání.

Běžně v teorii pravděpodobnosti je rozdělení pravděpodobnosti náhodné veličiny popsáno pomocí distribuční funkce. V analýze přežití k těmto účelům využíváme funkci přežití. Dále se k popisu náhodné veličiny T používá tzv. hazardní funkce, pro jejíž zavedení je nutné rozlišit, zda je náhodná veličina T spojitá nebo diskrétní.

A. Rozdělení pravděpodobnosti pro spojité T

Nechť T je spojitá, nezáporná náhodná veličina, jejíž rozdělení pravděpodobnosti je charakterizováno hustotou $f(t)$.

Definice 2.2. *Funkce přežití* spojité náhodné veličiny T je definována vztahem

$$S(t|\boldsymbol{x}) = P(T > t|\boldsymbol{x}), \qquad t \in [0, \infty]. \tag{2.1.1}$$

Funkce přežití při pevně daných regresorech $\boldsymbol{x}$, v časovém okamžiku t vyjadřuje podmíněnou pravděpodobnost toho, že u jedince z populace, která je určena regresory $\boldsymbol{x}$ nedojde k selhání v intervalu $[0, t]$. Jinými slovy, že se dožije času t.

Definice 2.3. *Hazardní funkce* $\lambda(t|\boldsymbol{x})$ spojité náhodné veličiny T je definována vztahem

$$\lambda(t|\boldsymbol{x}) = \lim_{h \to 0+} \frac{P(T \in [t, t+h]|T \geq t, \boldsymbol{x})}{h}, \qquad t \in [0, \infty]. \tag{2.1.2}$$

Hazardní funkce v čase t při pevně daných hodnotách regresorů $\boldsymbol{x}$, vyjadřuje podmíněnou pravděpodobnost selhání bezprostředně po čase t jedince z populace určené regresory $\boldsymbol{x}$, u nichž nedošlo k selhání nebo cenzorování v časovém intervalu $[0, t]$. Jinak řečeno

popisuje okamžité riziko selhání neselhavších a necenzorovaných jedinců dané populace v čase t.

Definice 2.4. *Kumulativní hazardní funkce* $\Lambda(t|\boldsymbol{x})$ spojité náhodné veličiny T je definována vztahem

$$\Lambda(t|\boldsymbol{x}) = \int_0^t \lambda(s|\boldsymbol{x})ds, \qquad t \in [0, \infty]. \tag{2.1.3}$$

Protože hazardní funkce je funkcí času, může někoho zajímat hazard jedince z dané populace za určitý časový interval. Na tuto otázku odpovídá kumulativní hazardní funkce.

Užitečné vztahy

Funkce přežití a distribuční funkce

$$S(t|\boldsymbol{x}) = 1 - F(t|\boldsymbol{x}). \tag{2.1.4}$$

Hazardní funkce a funkce přežití

$$\lambda(t|\boldsymbol{x}) = \frac{-\partial \log S(t|\boldsymbol{x})}{\partial t}. \tag{2.1.5}$$

Důkaz:

$$\lambda(t|\boldsymbol{x}) = \lim_{h\to 0^+} \frac{P(T \in [t, t+h]|\boldsymbol{x})}{P(T \geq t|\boldsymbol{x})h} = \lim_{h\to 0^+} \frac{P(T \geq t|\boldsymbol{x}) - P(T \geq t+h|\boldsymbol{x})}{P(T \geq t|\boldsymbol{x})h} =$$

$$= \lim_{h\to 0^+} \frac{S(t|\boldsymbol{x}) - S(t+h|\boldsymbol{x})}{S(t|\boldsymbol{x})h} = \frac{1}{S(t|\boldsymbol{x})} \lim_{h\to 0^+} \frac{F(t+h|\boldsymbol{x}) - F(t|\boldsymbol{x})}{h} = \frac{\frac{\partial F(t|\boldsymbol{x})}{\partial t}}{S(t|\boldsymbol{x})} =$$

$$= \frac{f(t|\boldsymbol{x})}{S(t|\boldsymbol{x})} = \frac{-\partial \log(S(t|\boldsymbol{x}))}{\partial t}$$

Funkce přežití a kumulativní hazardní funkce

$$S(t|\boldsymbol{x}) = exp\left(-\int_0^t \lambda(s|\boldsymbol{x})ds\right) = exp(-\Lambda(t|\boldsymbol{x})). \tag{2.1.6}$$

Důkaz:

$$-\int_0^t \lambda(s|\boldsymbol{x})ds = [\log(S(s|\boldsymbol{x}))]_0^t = \log(S(t|\boldsymbol{x}))-\log(S(0|\boldsymbol{x})) = \log(S(t|\boldsymbol{x}))$$

$$-\log(1) = \log(S(t|\boldsymbol{x})) \Rightarrow S(t|\boldsymbol{x}) = exp\left(-\int_0^t \lambda(s|\boldsymbol{x})ds\right) = exp(-\Lambda(t|\boldsymbol{x}))$$

Funkce hustoty a hazardní funkce

$$f(t|\boldsymbol{x}) = \lambda(t|\boldsymbol{x})exp(-\Lambda(t|\boldsymbol{x})). \tag{2.1.7}$$

B. Rozdělení pravděpodobnosti pro diskrétní T

Nechť T je diskrétní, nezáporná náhodná veličina nabývající hodnot $t_1 < t_2 < .....$, jejíž rozdělení pravděpodobnosti je charakterizováno pravděpodobnostní funkcí $p(t_j) = P(T = t_j),\ \ j = 1, 2, ...$

Definice 2.5. *Funkce přežití* diskrétní náhodné veličiny T je definována vztahem

$$S(t|\boldsymbol{x}) = P(T > t|\boldsymbol{x}). \tag{2.1.8}$$

Stejně jako u spoj. rozdělení T platí rovnost

$$S(t|\boldsymbol{x}) = 1 - F(t|\boldsymbol{x}).$$

Důkaz:

$$S(t|\boldsymbol{x}) = \sum_{j:t_j>t} p(t_j|\boldsymbol{x}) = 1 - \sum_{j:t_j\leq t} p(t_j|\boldsymbol{x}) = \ = 1 - F(t|\boldsymbol{x}), \quad t \in [0, \infty]$$

Definice 2.6. *Hazardní funkce* diskrétní náhodné veličiny T je definována pouze v bodech $t_j, j = 1, 2, ...$ vztahem

$$\begin{aligned} \lambda_j = \lambda(t_j|\boldsymbol{x}) = P(T = t_j|T \geq t_j, \boldsymbol{x}) = \frac{p(t_j|\boldsymbol{x})}{\lim\limits_{t\to t_j-} S(t|\boldsymbol{x})} = \\ = \frac{p(t_j|\boldsymbol{x})}{p(t_j|\boldsymbol{x}) + p(t_{j+1}|\boldsymbol{x}) + ...}. \end{aligned} \tag{2.1.9}$$

V ostatních bodech ji lze dodefinovat nulou, protože riziko selhání je mimo množinu $\{t_1, t_2, ...\}$ nulové.

Užitečné vztahy

Funkce přežití a hazardní funkce

$$S(t|\boldsymbol{x}) = \prod_{t_j\leq t} (1 - \lambda(t_j|\boldsymbol{x})). \tag{2.1.10}$$

Pravděpodobnostní funkce a hazardní funkce

$$p(t_j|\boldsymbol{x}) = \lambda(t_j|\boldsymbol{x}) \prod_{i=1}^{j-1} (1 - \lambda(t_i|\boldsymbol{x})). \tag{2.1.11}$$

2.2. Coxův model

Při analýze přežití v praxi pozorujeme realizace náhodného vektoru $(T, Y, \boldsymbol{x})$. Kde T představuje dobu přežití a Y je veličina, která říká, zda u jedince došlo k selhání či cenzorování. Pomocí nich pak odhadujeme závislost funkce přežití nebo hazardní funkce na regresorech $\boldsymbol{x} = (X_1, ..., X_p)$. Coxův model předpokládá hazardní funkci ve tvaru

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t) \exp(\boldsymbol{\beta}'\boldsymbol{x}), \tag{2.2.1}$$

kde $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ je vektor neznámých regresních koeficientů, $\boldsymbol{x} = (x_1, ..., x_p)'$ je vektor regresorů a $\lambda_0(t)$ neznámá, nezáporná základní hazardní funkce.

Z tvaru regresní funkce můžeme vyčíst předpoklady Coxova modelu. Jako první předpokládáme, že základní hazardní funkce nezávisí na hodnotách regresorů $\boldsymbol{x}$ a je tedy pro všechny jedince stejná. Za druhé, vektor regresních parametrů $\boldsymbol{\beta}$ nezávisí na čase. Stejně tak vektor regresorů $\boldsymbol{x}$ je nezávislý na čase, což je v případě naší studie splněno, kdy hodnoty regresorů pozorujeme pouze v jednom časovém okamžiku a to na začátku studie.

Předpoklad proporcionality Coxova modelu je nejčastěji vyjádřen prostřednictvím hazardního podílu. Kdy předpokládáme podíl hazardních funkcí konstantní v čase. Mějme dva jedince s funkcemi hazardu $\lambda(t|\boldsymbol{x_1})$ a $\lambda(t|\boldsymbol{x_2})$, pak podíl jejich hazardních funkcí

$$\frac{\lambda(t|\boldsymbol{x_1})}{\lambda(t|\boldsymbol{x_2})} = \frac{\lambda_0(t) \exp(\boldsymbol{\beta}'\boldsymbol{x_1})}{\lambda_0(t) \exp(\boldsymbol{\beta}'\boldsymbol{x_2})} = \exp(\boldsymbol{\beta}'(\boldsymbol{x_1} - \boldsymbol{x_2}))$$

je za splnění předchozích předpokladů v čase konstantní.

Interpretace regresních parametrů (*hazard ratio, HR*)

Poměr hazardních funkcí vyjadřuje efekt vysvětlujících proměnných na hazardní funkci. Popisuje, kolikrát se zvětší či zmenší hazard při posunu o jednotku u numerické proměnné nebo při přechodu z jedné kategorie do druhé u proměnných kategoriálních.

$$HR_i = \frac{\lambda(t|(x_1, ..., x_i + 1, ..., x_p)}{\lambda(t|x_1, ..., x_i, ..., x_p)} = \frac{\lambda_0(t) \exp\left(\sum_{j \neq i} \beta_j x_j + \beta_i(x_i + 1)\right)}{\lambda_0(t) \exp\left(\sum_{j=1}^{p} \beta_j x_j\right)} = \exp(\beta_i)$$

$$\Rightarrow \beta_i = \log\left(HR_i\right)$$

Pokud máme tedy dva jedince, kteří se liší pouze v hodnotě i-tého regresoru o jednotku, hodnota parametru β_i představuje logaritmus poměru hazardu. Pro lepší interpretaci se v Coxově modelu uvažují parametry ve tvaru,

$$HR_i = \exp(\beta_i)$$

který přímo vyjadřuje hazardní poměr.

V Coxově modelu platí následující vztahy

$$S(t|\boldsymbol{x}) = \exp\left(\int_0^t \lambda(s|\boldsymbol{x}ds)\right) = \exp^{-\exp \boldsymbol{\beta'x}\left(\int_0^t \lambda_0(s)ds\right)} \tag{2.2.2}$$

$$f(t|\boldsymbol{x}) = \lambda(t|\boldsymbol{x})S(t|\boldsymbol{x}) = \lambda_0(t) \exp(\boldsymbol{\beta'x} - \Lambda(t|\boldsymbol{x})) \tag{2.2.3}$$

$$S_0(t|\boldsymbol{x}) = \exp\left(-\int_0^t \lambda_0(s)ds\right), \quad \textit{tzv. základní funkce přežití} \tag{2.2.4}$$

$$f_0(t|\boldsymbol{x}) = \lambda_0(t|\boldsymbol{x})S_0(t|\boldsymbol{x}), \quad \textit{tzv. základní hustota} \tag{2.2.5}$$

Odhad parametrů modelu z dat - praktická ukázka

Zde uvádím odhadnutý model z datové sady uvedené v první kapitole. K určení nejlepšího modelu jsem použil *stepwise* algoritmus, kde jako určující kritérium jsem zvolil Bayesovo informační kritérium.

Poznámka 2.1 Bayesovo informační kritérium slouží k posouzení kvality regresního modelu. Pro Coxův model je Bayesovo informační kritérium určené na základě parciální věrohodnostní funkci (více v sekci 2.3.) a je dáno vztahem

$$BIC = -2\max_{\boldsymbol{\beta}}(L_p(\beta)) + p\log(n) = -2L_p(\widehat{\boldsymbol{\beta}}) + p\log(n),$$

kde L_p je logaritmus parciální věrohodnostní funkce, n je počet pozorování a p počet regresorů v modelu.

Stepwise algoritmus je metoda určená pro nalezení optimální volby regresorů do modelu. Optimální ve smyslu minimální hodnoty BIC. Algoritmus začíná s modelem se všemi regresory a v dalších krocích se na základě BIC rozhoduje, které regresory vyloučit nebo zahrnout do modelu. Algoritmus se zastaví ve chvíli, když jakákoliv eliminace regresoru zahrnutého v modelu, nebo přidání regresoru nezahrnutého v modelu vede k nárůstu BIC.

Kromě odhadů parametrů v tomto shrnutí uvádím také výsledky inference modelu, jako intervaly spolehlivosti nebo informaci o významnosti parametrů. Přičemž inferencí modelu se budu detailněji zabývat v další sekci. Pro další výpočty, jako např. diagnostiku modelu, budu využívat zde uvedený model.

Ukázka výstupu softwaru R:

```
BestCox <- coxph(Hazard ~ Věk+Tranfsúze+Stádium+Marker1+Marker4)

n= 103, number of events= 20

               coef   exp(coef) se(coef)      z     Pr(>|z|)
Věk           0.067   1.069     0.036       1.858   0.063 .
Transfúze     1.667   5.299     0.628       2.655   0.008 **
Stádium       0.931   2.538     0.205       4.543   5.54e-06 ***
Marker1       1.277   3.587     0.515       2.479   0.013 *
Marker4      -2.045   0.129     0.826      -2.475   0.013 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

             exp(coef)  exp(-coef)    lower .95    upper .95
Věk           1.069       0.936        0.996        1.147
Transfúze     5.299       0.189        1.547       18.143
Stádium       2.538       0.394        1.698        3.793
Marker1       3.587       0.279        1.307        9.845
Marker4       0.129       7.733        0.026        0.653

Likelihood ratio test  =   30.51  on 5 df,   p=1.168e-05
Wald test              =   25.79  on 5 df,   p=9.818e-05
Score (logrank) test   =   29.49  on 5 df,   p=1.861e-05
```

Regresor	$\exp \beta_i$	Konfidenční interval	p-hodnota
Věk	1.069	[0.996 ; 1.147]	0.063
Transfúze	5.299	[1.547 ; 18.143]	0.008
Stádium	2.538	[1.698 ; 3.793]	0.001<
Marker1	3.587	[1.307 ; 9.845]	0.013
Marker4	0.129	[0.026 ; 0.653]	0.013

Tabulka 2: Shrnutí odhadnutého modelu

Ve výše uvedené tabulce jsou stručně shrnuty důležité údaje odhadnutého modelu. V modelu je celkem 5 regresorů, které jsou dle posledního sloupce tabulky statisticky významné až na regresor `Věk` (ale `p-hodnota` je velmi blízko hladině významnosti $0,05$). Ve druhém sloupci můžeme vidět hodnotu, která říká, kolikrát se zvýší hazard při zvýšení hodnoty daného regresoru o jednotku. Zajímavý je odhad u regresoru `Marker4`. Ten říká, že přítomnost tohoto markeru v krvi snižuje hodnotu hazardu přibližně osmkrát. Ostatní regresory mají opačný efekt, kdy s jejich rostoucí hodnotou roste i hazard.

2.3. Inference v Coxově modelu

Pro odhad regresních parametrů a testování jejich významnosti, se v analýze přežívání používá metoda maximální věrohodnosti přizpůsobená pro cenzorovaná data.

Věrohodnostní funkce

Předpokládejme n nezávislých jedinců, kdy u každého z nich pozorujeme trojici $(t_i, y_i, \boldsymbol{x}_i), \quad i = 1, ..., n$. Kde t_i je čas, kdy došlo u i-tého jedince k selhání nebo cenzorování, y_i identifikuje příčinu ukončení pozorování u tohoto jedince. V případě, že došlo k selhání nabývá tato veličina hodnotu 1, v případě cenzorování hodnotu 0. Vektor $\boldsymbol{x}_i$ je vektor regresorů i-tého jedince.

Dále předpokládejme, že rozdělení pravděpodobností doby přežití i-tého jedince lze popsat distribuční funkcí $F(t, \boldsymbol{\beta}|\boldsymbol{x}_i)$ nebo funkcí hustoty $f(t, \boldsymbol{\beta}|\boldsymbol{x}_i)$. Pro konstrukci věrohodnostní funkce budeme uvažovat příspěvek trojic $(t_i, 1, \boldsymbol{x}_i)$ a $(t_i, 0, \boldsymbol{x}_i)$ odděleně. V případě trojice $(t_i, 1, \boldsymbol{x}_i)$ víme, že doba přežití je přesně t_i. Příspěvek této trojice k hodnotě věrohodností funkce je dán hodnotou funkce hustoty $f(t_i, \boldsymbol{\beta}, \boldsymbol{x}_i)$. Pro trojici $(t_i, 0, \boldsymbol{x}_i)$ víme, že doba přežití je nejméně t_i. Nyní je příspěvek pro věrohodnostní funkci dán jako pravděpodobnost, že doba přežití subjektu s regresory $\boldsymbol{x}_i$ je nejméně t_i. Tato pravděpodobnost je dána hodnotou funkce přežití $S(t_i, \boldsymbol{\beta}|\boldsymbol{x}_i)$. Potom úplný tvar věrohodnostní funkce je dán jako součin příslušných příspěvků pozorovaných trojic $(t_i, y_i, \boldsymbol{x}_i)$. Pro necenzorovaného hodnotou $f(t, \boldsymbol{\beta}|\boldsymbol{x})$ a pro cenzorovaného jedince hodnotou $S(t, \boldsymbol{\beta}|\boldsymbol{x})$. Obecně lze příspěvek i-tého jedince k hodnotě věrohodnostní funkce

zapsat ve tvaru

$$[f(t_i, \boldsymbol{\beta}|\boldsymbol{x}_i)]^{y_i} \times [S(t_i, \boldsymbol{\beta}|\boldsymbol{x}_i)]^{1-y_i},$$

kde $y_i = 0$ nebo 1.

Potom pro náš náhodný výběr n nezávislých jedinců dostaneme věrohodnostní funkci

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} [f(t_i, \boldsymbol{\beta}|\boldsymbol{x}_i)]^{y_i} \times [S(t_i, \boldsymbol{\beta}|\boldsymbol{x}_i)]^{1-y_i}. \tag{2.3.1}$$

Pro zjednodušení věrohodnostní funkce využijeme následující vztah, jeho platnost je ukázána v odvození vztahu (2.1.5) a dostaneme

$$f(t|\boldsymbol{x}) = \lambda(t|\boldsymbol{x})S(t|\boldsymbol{x}).$$

Po dosazení za funkci hustoty dostaneme

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} [\lambda(t_i, \boldsymbol{\beta}|\boldsymbol{x}_i)S(t_i, \boldsymbol{\beta}|\boldsymbol{x}_i)]^{y_i} \times [S(t_i, \boldsymbol{\beta}|\boldsymbol{x}_i)]^{1-y_i}. \tag{2.3.2}$$

Pro účel maximalizace věrohodnostní funkce vzhledem k parametru $\boldsymbol{\beta}$ budeme uvažovat (z důvodu menší výpočetní náročnosti) logaritmus věrohodnostní funkce

$$\begin{aligned} L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} y_i \log [\lambda(t_i, \boldsymbol{\beta}|\boldsymbol{x}_i)S(t_i, \boldsymbol{\beta}|\boldsymbol{x}_i)] \times (1 - y_i) \log [S(t_i, \boldsymbol{\beta}|\boldsymbol{x}_i)] = \\ &= \sum_{i=1}^{n} y_i \log [\lambda(t_i, \boldsymbol{\beta}|\boldsymbol{x}_i)] + \log [S(t_i, \boldsymbol{\beta}|\boldsymbol{x}_i)]. \end{aligned} \tag{2.3.3}$$

Po dosazení vztahů 2.2.1 a 2.2.2 do rovnice 2.3.3 dostaneme

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log(\lambda_0(t_i)\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)) + \log\left(\exp^{\left(-\int_0^t \lambda_0(s)ds\right)\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}\right) =$$

$$= \sum_{i=1}^{n} y_i \log(\lambda_0(t_i)\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)) + \log\left(S_0(t_i|\boldsymbol{x}_i)^{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}\right) =$$

$$= \sum_{i=1}^{n} y_i \log(\lambda_0(t_i)) + y_i\boldsymbol{\beta}'\boldsymbol{x}_i + \log(S_0(t_i|\boldsymbol{x}_i))\exp(\boldsymbol{\beta}'\boldsymbol{x}_i). \tag{2.3.4}$$

Derivace logaritmu věrohodnostní funkce podle parametru $\boldsymbol{\beta}$ má tvar

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} y_i\boldsymbol{x}_i + \boldsymbol{\beta}'\boldsymbol{x}_i \log(S_0(t_i|\boldsymbol{x}_i))\exp(\boldsymbol{\beta}'\boldsymbol{x}_i). \tag{2.3.5}$$

Pro určení maximálně věrohodného odhadu $\boldsymbol{\beta}$ je nutné vyřešit soustavu rovnic

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} \stackrel{!}{=} 0 \quad j = 1, ..., p.$$

Pro řešení této soustavy je nutné znát základní hazardní funkci $\lambda_0(t_i)$, ta je však z definice modelu neznámá (v podstatě se jedná o neznámý parametr modelu, který však nechceme odhadovat) a proto není možné určit maximálně věrohodný odhad parametru $\boldsymbol{\beta}$.

Parciální věrohodnostní funkce

Tento problém odhadu regresních parametrů $\boldsymbol{\beta}$ vyřešil *Cox*, který navrhl využít tzv. *parciální věrohodnostní funkci*, která závisí pouze na parametrech, které chceme odhadovat. *Cox* ve svém přístupu předpokládal, že maximálně parciálně věrohodný odhad by měl mít stejné distribuční vlastnosti jako maximálně věrohodný odhad (rigorózní matematický důkaz založený na teorii martingalů je uveden v [**?**], kapitola7).

Hlavní myšlenkou, která stojí za konstrukcí parciální věrohodnostní funkce, je to, že časové okamžiky, ve kterých nedošlo k selhání, nenesou žádnou informaci o efektu regresorů na hazard selhání. Lze totiž předpokládat, že v čase, kdy nedošlo k selhání, je základní hazardní funkce $\lambda_0(t)$ nulová a tedy i hazardní funkce $\lambda(t)$ je v tomto čase nulová. Z toho plyne, že informaci o efektu regresorů na hazardní funkci nesou pouze časy selhání.

Mějme množinu indexů D obsahující indexy těch jedinců, u nichž došlo k selhání. Dále uvažujme pravděpodobnost jevu selhání i-tého jedince v čase t_i, podmíněnou jevem $i \in D$ (v čase t_i došlo k selhání). Jestliže vektor regresorů jedince selhavšího v čase t_i označíme jako $\boldsymbol{x}_i$, pak tuto pravděpodobnost vyjádříme následovně

$$\text{P(selhání jedince s regersory } \boldsymbol{x}_i \text{ v čase } t_i | \text{v čase } t_i \text{ došlo k selhání právě jednoho jedince}) =$$

$$= \frac{\text{P(selhání jedince s regersory } \boldsymbol{x}_i \text{ v čase } t_i)}{\text{P(v čase } t_i \text{ došlo k selhání jednoho jedince)}}.$$

Pak tuto podmíněnou pravděpodobnost můžeme vyjádřit pomocí hazardní funkce jako podíl, kde v čitateli je pouze hazardní funkce v čase t_i jedince, jehož vektor regresorů je $\boldsymbol{x}_i$ a ve jmenovateli je suma

hodnot hazardní funkce v čase t_i všech jedinců, kteří jsou v tomto čase ohrožení selháním. Označme tuto množinu jedinců ohrožených v čase t_i jako R_i:

$$\frac{\lambda(t_i|\boldsymbol{x}_i)}{\sum\limits_{k\in R_i}\lambda(t_i|\boldsymbol{x}_k)} = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}{\sum\limits_{k\in R_i}\exp(\boldsymbol{\beta}'\boldsymbol{x}_k)}. \tag{2.3.6}$$

Jak už bylo řečeno, informaci o efektu regresorů na hazardní funkci máme pouze v časech selhání. Proto je parciální věrohodnostní funkce definována jako součin podmíněných pravděpodobností (2.3.5) v časech (selhání) t_i s indexem $i \in D$

$$l_p(\boldsymbol{\beta}) = \prod_{i\in D} P(\boldsymbol{X}_i = \boldsymbol{x}_i|R_i) = \prod_{i\in D}\frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}{\sum\limits_{k\in R_i}\exp(\boldsymbol{\beta}'\boldsymbol{x}_k)}. \tag{2.3.7}$$

Potom logaritmus parciální věrohodnostní funkce má tvar

$$L_p(\boldsymbol{\beta}) = \sum_{i\in D}\left(\boldsymbol{\beta}'\boldsymbol{x}_i - \log\left(\sum_{k\in R_i}\exp(\boldsymbol{\beta}'\boldsymbol{x}_k)\right)\right). \tag{2.3.8}$$

Derivací logaritmu parciální věrohodnostní funkce dostaneme

$$\frac{\partial L_p(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \sum_{i\in D}\left(\boldsymbol{x}_i - \frac{\sum\limits_{k\in R_i}\exp(\boldsymbol{\beta}'\boldsymbol{x}_k)}{\sum\limits_{k\in R_i}\exp(\boldsymbol{\beta}'\boldsymbol{x}_k)}\right). \tag{2.3.9}$$

Řešením soustavy rovnic

$$\sum_{i\in D}\left(x_{ij} - \frac{\sum\limits_{k\in R_i}\exp(\beta_j' x_{kj})}{\sum\limits_{k\in R_i}\exp(\beta_j' x_{kj})}\right) \stackrel{!}{=} 0 \quad j = 1, ..., p \tag{2.3.10}$$

určíme maximálně parciálně věrohodný odhad regresních parametrů $\widehat{\boldsymbol{\beta}}$.

Zajímat nás také bude odhad rozptylu maximálně parciálně věrohodného odhadu $\widehat{\boldsymbol{\beta}}$. Tento odhad je určen způsobem, jak je to v teorii maximální věrohodnosti nejběžnější. Odhad je inverzí záporu druhé derivace logaritmické parciální věrohodnostní funkce vzhledem k parametru $\boldsymbol{\beta}$.

Druhá derivace logaritmu parciální věrohodnostní funkce vzhledem k parametru $\boldsymbol{\beta}$ má tvar

$$\frac{\partial^2 L_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = -\sum_{i \in D} \frac{\left(\sum\limits_{k \in R_i} \exp(\boldsymbol{\beta}' \boldsymbol{x}_k)\right)\left(\sum\limits_{k \in R_i} x_k^2 \exp(\boldsymbol{\beta}' \boldsymbol{x}_k\right)}{\sum\limits_{k \in R_i} \exp(\boldsymbol{\beta}' \boldsymbol{x}_k)} + \frac{\left(\sum\limits_{k \in R_i} x_k \exp(\boldsymbol{\beta}' \boldsymbol{x}_k)\right)^2}{\sum\limits_{k \in R_i} \exp(\boldsymbol{\beta}' \boldsymbol{x}_k)} \quad . \tag{2.3.11}$$

Záporná hodnota logaritmu parciální věrohodnostní funkce se nazývá *výběrová informační matice* a značíme ji jako

$$I(\boldsymbol{\beta}) = -\frac{\partial^2 L_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}, \tag{2.3.12}$$

kde $I(\boldsymbol{\beta})$ je čtvercová matice typu $p \times p$.

Potom odhad varianční matice maximálně parciálně věrohodného odhadu je dán vztahem

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = I(\widehat{\boldsymbol{\beta}})^{-1}. \tag{2.3.13}$$

Testy signifikance

(1) `Waldův test`

Klasický Waldův test pro testování významnosti regresních parametrů, se užívá i v případě Coxova modelu. Test můžeme využít pro testování významnosti všech nebo pouze několika regresních parametrů. Nechť $\boldsymbol{\beta}_0$ obsahuje q nenulových, neznámých parametrů. Testujeme hypotézu

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \text{ proti alternativě } H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0,$$

pak

$$Z = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' I(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

Z má za platnosti nulové hypotézy asymptoticky chí-kvadrát rozdělení o $p - q$ stupních volnosti.

(2) `Test založený na poměru parciálních věrohodnostních funkcí`

Testová statistika založena na poměru parciálních věrohodnostních funkcí. Test můžeme využít pro testování významnosti všech nebo pouze několika regresních parametrů. Nechť $\boldsymbol{\beta}_0$ obsahuje q nenulových, neznámých parametrů. Testujeme hypotézu

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \text{ proti alternativě } H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0,$$

pak

$$G = 2[l_p(\widehat{\boldsymbol{\beta}}) - l_p(\boldsymbol{\beta}_0)]$$

G má za platnosti nulové hypotézy asymptoticky chí-kvadrát rozdělení o $p - q$ stupních volnosti.

3. Diagnostika Coxova modelu

Stejně jako v tradičním lineárním regresním modelu, tak i v Coxově modelu se k diagnostice využívají rezidua. Avšak oproti tradičnímu modelu, kde pracujeme s reziduii (případně jejich transformací), která jsou dána vztahem $\widehat{\epsilon_i} = Y_i - f(x_i, \widehat{\boldsymbol{\beta}})$, existuje pro diagnostiku Coxova modelu existuje několik typů reziduií, přičemž jejich výpočet a další použití je značně náročnější. V první řadě se zaměříme na Schoenfeldova rezidua, která jsou klíčová pro test předpokladu proporcionality modelu a podíváme se, jakými způsoby lze tento předpoklad ověřit. Dále se podíváme na diagnostiku pomocí martingalových reziduí, která je doplněna simulační studií. V této kapitole je čerpáno z [**?**], [**?**], [**?**] a [**?**]

3.1. Schoenfeldova rezidua

Prvním typem reziduí jsou tzv. Schoenfeldova rezidua, která slouží k ověření předpokladu proporcionality hazardu. Netradiční vlastností těchto reziduií je to, že je neodhadujeme pro všechna pozorování, ale pouze pro ta, u nichž pozorujeme selhání.

Konstrukce Schoenfeldových reziduí

Konstrukce a odhad Schoenfeldových reziduí vychází přímo z úvah pro konstrukci parciální věrohodnostní funkce.
Pro pevné $i \in D$ vybereme libovolný subjekt u něhož existuje v čase t_i riziko selhání z R_i. Index tohoto subjektu označme m. Za podmínky, že u právě jednoho ze subjektů s indexy v R_i dojde v čase t_i k události, bude to subjekt s indexem m s pravděpodobností

$$P(Z_i = m) = \exp(\boldsymbol{\beta}' \boldsymbol{X}_m) / \sum_{k \in R_i} \exp(\boldsymbol{\beta}' \boldsymbol{X}_k), \tag{3.1.1}$$

kde Z_i je náhodná veličina představující index selhavšího v čase t_i. Nyní $\boldsymbol{X}_i$ považujeme za náhodný vektor. Pro střední hodnotu j-té kovariáty selhavšího v čase t_i platí vztah

$$E(X_{ij}|R_i) = \sum_{k \in R_i} X_{kj} P(Z_i = k) = \frac{\sum\limits_{k \in R_i} X_{kj} \exp(\boldsymbol{\beta}' \boldsymbol{X}_k)}{\sum\limits_{k \in R_i} \exp(\boldsymbol{\beta}' \boldsymbol{X}_k)} \qquad (3.1.2)$$

Regresní parametr $\boldsymbol{\beta}$ odhadujeme pomocí metody maximální parciální věrohodnosti. Řešením systému věrohodnostních rovnic určíme odhad parametru $\boldsymbol{\beta}$

$$\frac{\partial \log(L(\boldsymbol{\beta}))}{d\beta_j} = \sum_{i \in D} \left(X_{ij} - \frac{\sum\limits_{k \in R_i} X_{kj} \exp(\boldsymbol{\beta}' \boldsymbol{x}_k)}{\sum\limits_{k \in R_i} exp(\boldsymbol{\beta}' \boldsymbol{x}_k)} \right) =$$

$$= \sum_{i \in D} (X_{ij} - E(X_{ij}|R_i)) = 0, \quad j = 1, ..., p.$$

Poznámka 3.1. Je dobré si všimnout, že odhad parametru $\widehat{\boldsymbol{\beta}}$ nezávisí pouze na hodnotách jedinců, u nichž došlo k selhání, neboť rozdělení pravděpodobnosti náhodného vektoru kovariátů $\boldsymbol{X}$ selhavších jedinců je podmíněné tím, kteří jedinci jsou stále v ohrožení selhání v době, kdy u daného selhavšího jedince dochází k selhání.

Definice 3.2. (Schoenfeldova rezidua) Řešením soustavy věrohodnostních rovnic obdržíme maximálně věrohodný odhad $\widehat{\boldsymbol{\beta}}$, po dosazení tohoto odhadu do rovnice 3.1.2. dostaneme odhad $\widehat{E}(X_{ij}|R_i)$. Nechť t_i značí čas selhání i-tého selhavšího subjektu z D, pak vektor Schoenfeldových reziduí tohoto jedince v tomto čase je vektor $\widehat{\boldsymbol{r}}_i = (\widehat{r}_{i1}, ..., \widehat{r}_{ip})'$, daný vztahem

$$\widehat{r}_{ik} = X_{ik} - \widehat{E}(X_{ik}|R_i), \quad k = 1, ..., p. \qquad (3.1.3)$$

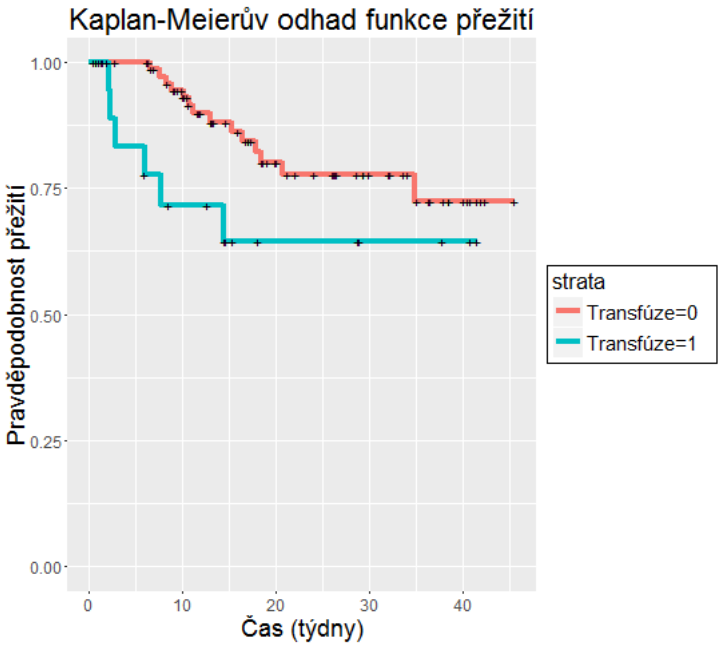
3.2. Testování proporcionality hazardu

V této sekci si představíme možnosti ověření předpokladu proporcionality Coxova modelu. Na úvod si představíme jednoduchou neexaktní grafickou metodu, která je založena na Kaplan-Meierově odhadu funkce přežití. Dále si představíme třídu testů pro testování předpokladu proporcionality, které doplníme i grafickým přístupem.

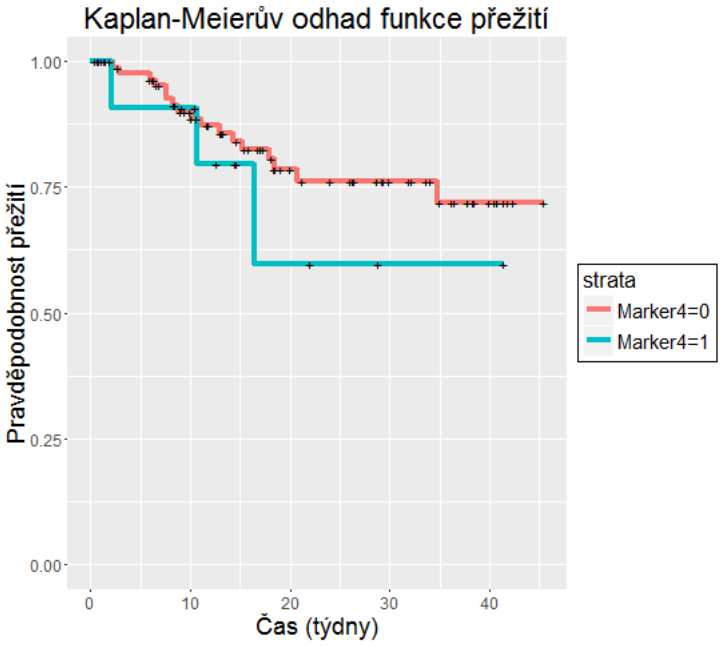
3.2.1. Kaplan-Meierovův odhad funkce přežití

První představu o předpokladu proporcionality hazardu můžeme dostat zobrazením Kaplan-Meierova odhadu funkce přežití. Na základě tohoto přístupu není možné rigorózně rozhodnout, zda předpoklad je splněn či ne, ale pouze získat základní informaci. Navíc je tento přístup limitován počtem a typem proměnných, které do modelu vstupují. Nelze jím např. ověřovat předpoklad pro spojité proměnné. Tento přístup je vhodný pro malý počet proměnných, které nabývají jen několika málo hodnot.

Zajímá nás, jestli křivky funkcí přežití jednotlivých populací jsou v čase proporcionální.



Obrázek 1: Funkce přežití dvou populací, které jsou určené logickou proměnnou Transfúze



Obrázek 2: Funkce přežití dvou populací, které jsou určené logickou proměnnou Marker4

Na obrázku 1 můžeme vidět, že průběh křivek funkcí přežití je podobný. V tomto případě tedy nemáme důvod předpokládat, že náš předpoklad byl porušen. Ale na obrázku 2 můžeme vidět, že dochází k dvojímu protnutí křivek funkcí přežití a zde bychom mohli pochybovat o platnosti našeho předpokladu.

3.2.2. Grambschové-Therneauova třída testů

Grambschová a Thernau navrhli test založený na normovaných Schoenfeldových reziduích, který testuje předpoklad proporcionality pro jednotlivé regresory v modelu. Použití normovaných Schoenfeldových reziduí, vede k větší diagnostické síle než použití obyčejných reziduí.

Definice 3.3. (normovaných Schoenfeldových reziduí)

Mějmě vektor Schoenfeldových reziduí pro i-tého jedince z množiny D dán jako $\widehat{\boldsymbol{r}}_i = (\widehat{r}_{i1}, ..., \widehat{r}_{ip})$.

Dále označme odhad $p \times p$ kovarianční matice tohoto vektoru jako $\widehat{Var}(\widehat{\boldsymbol{r}}_i)$.

Potom vektor normovaných Schoenfeldových reziduí i-tého jedince z množiny D je součinem inverze odhadnuté kovarianční matice a vektoru Schoenfeldových reziduí

$$\widehat{\boldsymbol{r}}_i^{\star} = \left[\widehat{Var}(\widehat{\boldsymbol{r}}_i)\right]^{-1} \widehat{\boldsymbol{r}}_i. \tag{3.2.1}$$

Prvky kovarianční matice $\widehat{Var}(\widehat{\boldsymbol{r}}_i)$ jsou, váženou verzí obvyklé sumy čtverců počítané na množině jedinců ohrožených selháním v čase t_i.

Pro i-tého jedince, diagonální prvky této matice mají tvar (dle [**?**], strana 199)

$$\widehat{Var}(\widehat{\boldsymbol{r}}_i)_{kk} = \sum_{j \in R_i} w_{ij}(X_{jk} - \widehat{E}(X_{jk}|R_i))^2, \tag{3.2.2}$$

a mimodiagonální prvky jsou dány jako (dle [**?**], strana 200)

$$\widehat{Var}(\widehat{\boldsymbol{r}}_i)_{kl} = \sum_{j \in R_i} w_{ij}(X_{jk} - \widehat{E}(X_{jk}|R_i))(X_{jl} - \widehat{E}(X_{jl}|R_i)), \tag{3.2.3}$$

kde

$$w_{ij} = \frac{\exp \boldsymbol{\beta}' \boldsymbol{X}_j}{\sum\limits_{l \in R_i} \exp \boldsymbol{\beta}' \boldsymbol{X}_l}.$$

Pro jednoduší výpočet byla navrhnuta aproximace normovaných Schoenfeldových reziduií, která je založena na zkušenosti, že matice $\widehat{Var}(\widehat{\boldsymbol{r}}_i)$ má tendenci chovat se v čase konstantně. Pokud je tato matice konstantní, její inverze může být aproximována kovarianční maticí odhadnu $\widehat{\boldsymbol{\beta}}$ regresního parametru $\boldsymbol{\beta}$, přenásobenou počtem selhání d,

$$\left[\widehat{Var}(\widehat{\boldsymbol{r}}_i)\right]^{-1} \approx d \cdot \widehat{Var}(\widehat{\boldsymbol{\beta}}). \tag{3.2.4}$$

Aproximaci normovaných Schoenfeldových reziduí pak dostaneme jako

$$\widehat{\boldsymbol{r}}_i^{\star} \approx d \cdot \widehat{Var}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{r}}_i. \tag{3.2.5}$$

Definice 3.4. (Testová statistika) Existuje velké množství způsobů, jak modelovat a testovat neproporcionalitu v Coxově modelu. Nicméně Grambschová a Thernau ukázali, že existuje jednoduchý test a na něm založená grafická metoda, které jsou účinným nástrojem pro ověření proporcionality hazardu Coxova modelu.

Z rovnice 2.2.1 určíme logaritmus hazardní funkce

$$\log(\lambda(t|x)) = \log(\lambda_0(t)) + \boldsymbol{\beta}'\boldsymbol{x},$$

kdy autoři testu navrhují porušit předpoklad proporcionality modelu následovně:

Místo konstantních regresních parametrů uvažujeme regresní parametry závislé na čase, kdy tato závislost je popsána vztahem

$$\beta_k(t) = \beta_k + \gamma_k g_k(t), \quad k = 1, ..., p \tag{3.2.6}$$

kde $g_k(t)$ nějaká (daná) transformace času.

Za platnosti modelu 3.2.6 pro normovaná Schoenfeldova rezidua (3.2.1) a jejich aproximaci (3.2.5) pro k-tý regresor platí vztah

$$E[r_k(t)] \approx \gamma_k g_k(t). \tag{3.2.7}$$

Tento vztah je velmi důležitý, protože je na něm založena grafická metoda pro ověření proporcionality. Graf závislosti normovaných Schoenfeldových reziduí na čase můžeme použít k úvaze o nulovosti γ_k. Pokud předpoklad proporcionality neplatí, můžeme získat informací o funkční závislosti parametru β_k na čase vyjádřené pomocí funkce $g_k(t)$. Pro test o nulovosti γ_k byl odvozen zobecněný odhad tohoto koeficientu metodou nejmenších čtverců a skórová statistika pro hypotézu o nulovosti tohoto koeficientu, při dané transformaci času $g_k(t)$.

Při testování hypotézy

$$H_0 : \gamma_k = 0 \text{ proti } H_1 : \gamma_k \neq 0,$$

má testová statistika pro k-tý regresor tvar

$$T_k = \frac{\left(\sum_{i \in D} (g_k(t_i) - \bar{g_k}) \widehat{r}^{\star}_{ik} \right)^2}{d \cdot \widehat{Var}(\widehat{\beta}_k) \sum_{i \in D} (g_k(t_i) - \bar{g_k})^2},$$

kde

$$\bar{g_k} = \frac{1}{|D|} \sum_{i \in D} g_k(t_i).$$

Funkce g zde představuje transformaci času. Nejčastěji používané transformace jsou identita nebo přirozený logaritmus. Tato transformace určuje Grambschové-Thernauovu třídu testů. Tyto testy jsou implementovány v knihovně *survival* softwaru *R* ve funkci `cox.zph`. Argument `transfrom` této funkce specifikující časovou transformaci, umožňuje volbu čtyř různých transformací, které pak určují název testů z této třídy.

1. **"identity"** - test pracuje s původní časem selhání a nazývá se Coxův test
2. **"log"** - test pracuje s logaritmem času selhání a nazývá se Coxův test
3. **"rank"** - test pracuje s pořadím času selhání a nazývá se Breslowův-Edlerův-Bergerův test
4. **"km"** - test pro transformaci času využívá Kaplan-Meierův odhad funkce přežití a nazývá se Linův test

Testová statistika má pak za platnosti nulové hypotézy chí-kvadrát rozdělení o jednom stupni volnosti a může být interpretována jako míra korelace mezi normovanými Schoenfeldovými rezidui určitého regresoru a časem selhání. Pokud testujeme na hladině významnosti α, pak proporcionalitu k-tého regresoru zamítáme pokud T_k je větší než $(1 - \alpha)$ kvantil chí-kvadrát rozdělení o jednim stupni volnosti.

Grambschové-Therneaův test na reálných datech

Nyní provedeme ověření předpokladu proporcionality na námi odhadnutém modelu pomocí testu a graficky.

```
Model: Hazard ~ Věk+Transfúze+Stádium+Marker1+Marker4

cox.zph(BestCox,transform="identity")
                rho    chisq     p
Věk          -0.106   0.239   0.625
Transfúze    -0.320   2.934   0.087
Stádium       0.163   0.384   0.536
Marker1       0.092   0.174   0.677
Marker4       0.051   0.086   0.770
GLOBAL           NA   5.238   0.388
```

Sloupec *rho* představuje korelační koeficient mezi normovanými Schoenfeldovými rezidui určitého regresoru a časem selhání, sloupec *chisq* představuje hodnotu testové statistiky a sloupec p je p-hodnota testu. Můžeme vidět, že pokud pro transformaci času uvažujeme identitu, pak žádný regresor signifikantně neporušuje předpoklad proporcionality na hladině významnosti $0,05$. I když regresor Transfusion se jeví problematicky.

Ovšem pokud místo identity budeme uvažovat logaritmickou závislost, dostaneme poněkud odlišné výsledky:

```
Model: Hazard ~ Věk+Transfúze+Stádium+Marker1+Marker4

cox.zph(BestCox,transform="log")
               rho    chisq     p
Věk         -0.092   0.179   0.672
Transfúze   -0.436   5.452   0.020
Stádium     -0.049   0.035   0.852
Marker1      0.129   0.339   0.560
Marker4      0.095   0.293   0.588
GLOBAL          NA   7.065   0.216
```

U regresoru Transfúze test detekuje porušení proporcionality.

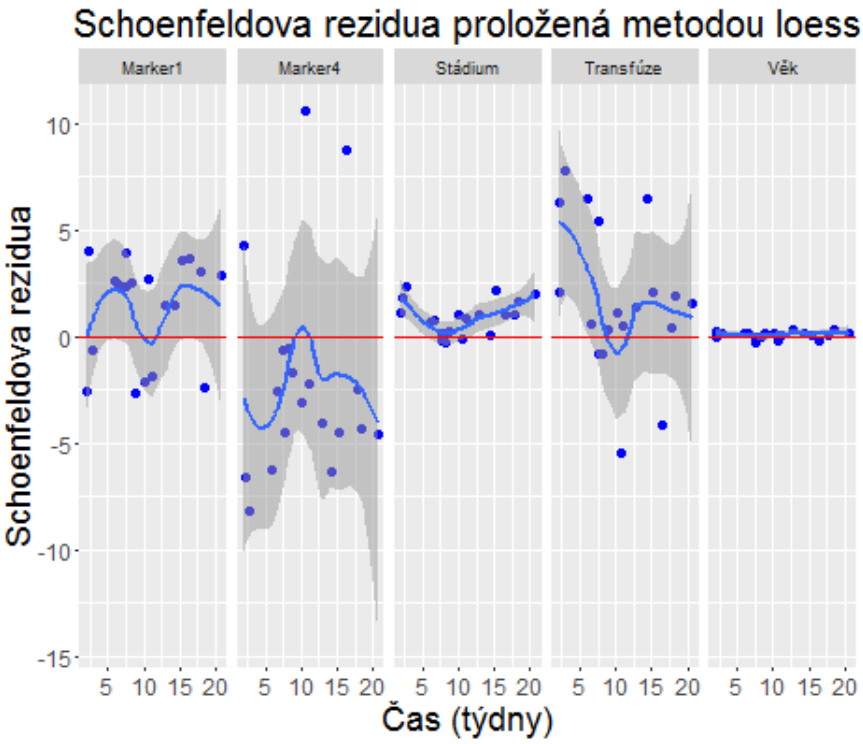
Grambschové-Therneaův test - grafický přístup

Testování, zda se v modelu vyskytují časově závislé regresory, je ekvivalentní tomu, že testujeme, zda závislosti normovaných Schoenfeldových reziduí na čase je nulová. Nenulová směrnice je indikátorem porušení předpokladu proporcionality hazardu. Jako u běžné regrese i zde se doporučuje samotný test podpořit graficky. Existují typy neproporcionality, které nebudou odhaleny testem nulovosti směrnice, ale mohou být dobře viditelné z grafu, např. nelineární vztah mezi normovanými Schoenfeldovými rezidui a časem. Problémem mohou být i odlehlá pozorování. Pro orientaci můžeme test zopakovat s různým nastavením transformace času a podívat se , jestli z grafů nelze rozpoznat časovou závislost.

Návíc díky vztahu (3.2.7) můžeme hodnoty normovaných Schoenfeldových reziduí považovat za hodnoty daného regresního koeficientu v čase. To je také důvod, proč v softwaru `R` výsledný graf funkce

`cox.zph` na osu `y` vynáší hodnoty β_k. My tedy pozorujeme vývoj daného regresního parametru v čase a můžeme na základě této grafické informace usuzovat, zda se v čase chová konstantně nebo se mění na základě určité funkční závislosti.

Modrá křivka v grafu reziduí je do grafu přidána metodou *loess* (Local Polynomial Regression Fitting). Proložení je provedeno lokálně. Pro proložení bodu t pracujeme s body z okolí bodu t, které vážíme jejich vzdáleností od bodu t. Velikost okolí bodu t určíme ladícím parametrem α. Defaultní nastavení regresní metody je metoda vážených nejmenších čtverců. Šedě znázorněná oblast kolem regresní křivky je pás spolehlivosti. V našem rezidua prokládáme lineární regresní funkcí.



Obrázek 3: Schoenfeldova rezidua jednotlivých regresorů v čase, lokálně proložené lineární funkcí

3.3. Martingalová rezidua

Martingalová rezidua slouží k určení funkční závislosti hazardní funkce na spojitých regresorech, které vystupují v modelu. Někdy se také v literatuře uvádí, že slouží k ověření tzv. předpokladu linearity spojitých regresorů v Coxově modelu. Tento předpoklad vychází z toho, že v exponentu linkové funkce máme lineární kombinaci všech regresorů. Přičemž je vhodné poznamenat, že Coxův model primárně nepředpokládá, že všechny spojité regresory mají v exponentu linkové funkce lineární vliv na hazardní funkci. Pouze v exponentu linkové funkce předpokládá lineární kombinaci $(\beta_1 x_1 + ... + \beta_p x_p)$, přičemž místo regresoru x_k můžeme uvažovat jeho libovolnou funkci $f_k(x_k)$, kde $k = 1, ..., p$.

Definice 3.5. (Martingalová rezidua) Martingalová rezidua jsou na rozdíl od těch Schoenfeldových odhadnuta pro všechny jedince ve studii. Pro i-tého jedince definujeme martingalové reziduum vztahem

$$\widehat{M_i} = y_i - \widehat{\Lambda}(t_i|\boldsymbol{x_i}), \tag{3.3.1}$$

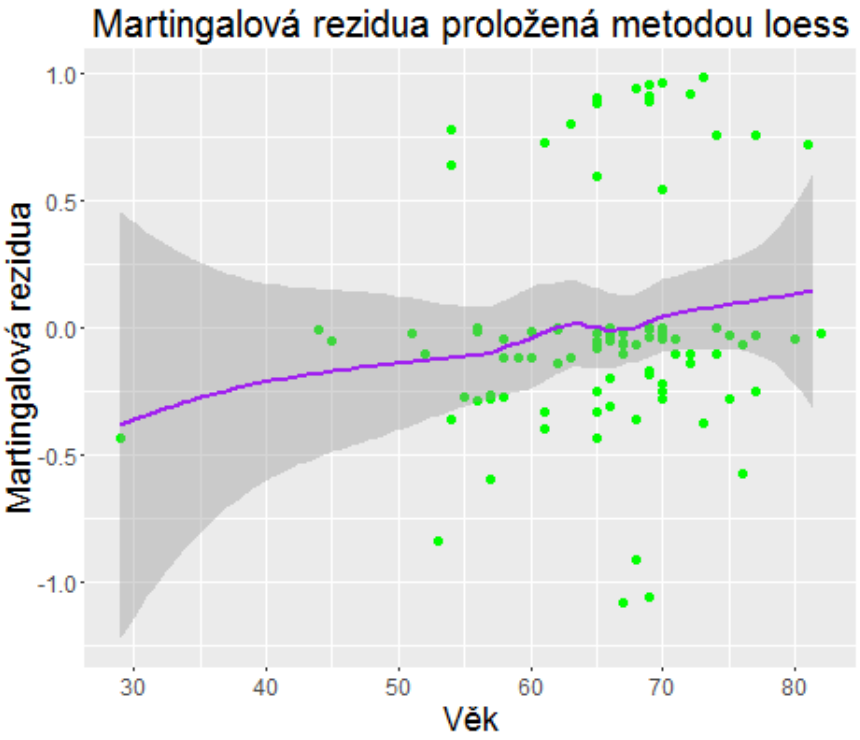
kde y_i je proměnná (kódovaná $0-1$) popisující, zda u i-tého jedince došlo k selhání.

Martingalová rezidua nabývají hodnot z intervalu $[-\infty, 1]$ a lze je intepretovat jako rozdíl pozorovaného selhání u jednotlivých osob ve studii a jejich „očekávaný“ počet selhání za dobu, po kterou daný jedinec je zahrnut do studie, založený na odhadnutém modelu.

Pro diagnostiku pomocí martingalových reziduí se uplatňuje následující postup: Spojitý regresor, jehož funkční vliv na funkci hazardu

chceme prověřit, z modelu vypustíme. Odhadneme tento nový model a pomocí něj určíme martingalová rezidua. Zkonstruujeme graf závislosti těchto reziduí na daném regresoru. Body grafu proložíme křivkou tak, aby co nejlépe vystihovala tento funkční vztah. K tomu existuje celá řada metod, my stejně jako u diagnostiky Schoenfeldových reziduí budeme pracovat s metodou *loess*.

V modelu odhadnutém v druhé kapitole máme pouze jeden spojitý regresor a to věk pacienta.



Obrázek 4: Ověření funkčního vlivu proměnné Věk na funkci hazardu

Můžeme vidět, že do 50 let ve studii nemáme téměř žádné pozorování a od 50 let je vliv věku lineární s kladnou směrnicí, což odpovídá odhadu regresního koeficienu pro věk.

3.3.1. Simulace funkční závislosti hazardní funkce na spojitém regresoru

V této sekci jsem se rozhodl prověřit diagnostické vlastnosti martingalových reziduí. Pro tento účel vytvořím simulaci, kde budu simulovat funkční závislost hazardní funkce na spojitém regresoru. Pro diagnostiku této závislosti použiji způsob zmíněný již v kapitole 3.3.

Nastavení simulace
Pro jednoduchost uvažuji pouze jeden regresor spojitého typu x.

- Regresor x nabývá hodnot na intervalu $[-1, 1]$. Pro simulaci volím 10000 ekvidistantně rozdělených hodnot z tohoto intervalu.
- Parametr D popisuje délku studie. Moje volba je D=10 (např. 10 týdnů).
- Protože pro realizaci simulace je nutné pracovat s diskrétním časem, budeme potřebovat parametr h pro diskretizaci času. Ten určuje vzdálenost okamžiků, kdy mohu pozorovat selhání. Zvolil jsem $h = 0.01$, tím dostáváme $D/h = 1000$ potencionálních časových okamžiků selhání pro každého jedince.
- Parametr $\beta = \ln(3)$, touto volbou říkám, že jedinec je s regresorem rovným jedné má třikrát vyšší riziko selhání než jedinec s regresorem rovným nule.
- Základní hazardní funkci volím $\lambda_0(t) = 0.05$.

Nyní ze vztahu

$$\lambda(t|x) = \lim_{h \to 0^+} \frac{P(T \in [t, t+h]|T \geq t, x)}{h},$$

při konstantní volbě $\lambda_0(t)$ dostaneme:

$$\lambda(x) = \lim_{h \to 0^+} \frac{P(T \in [t, t+h]|x)}{hP(T \geq t|x)}$$

$$\Rightarrow P(T \in [t, t+h]|x) \approx \lambda(x)P(T \geq t)h.$$

V prvním časovém intervalu dostaneme

$$P(T \in [0, h]|x) \approx \lambda(x)P(T \geq 0)h = \lambda(x)h.$$

V druhém časovém intervalu dostaneme

$$P(T \in [h, 2h]|x) \approx \lambda(x)P(T \geq h)h = \lambda(x)h - \lambda^2(x)h^2.$$

Tyto vztahy říkají, jaká je pravděpodobnost selhání jedince s daným regresorem x na jednom z $\frac{D}{h}$ časových intervalů, kde každý z těchto intervalů má délku h. Pravděpodobnost přežití přirozeně v čase neroste. To, že dojde k selhání na intervalu $[0, h]$ je pravděpodobnější, než na intervalu $[h, 2h]$. Nás ovšem zajímá, jaká je podmíněná pravděpodobnost selhání na intervalu $[0, h]$ za podmínky dožití se časového okamžiku 0, respektive selhání na intervalu $[h, 2h]$ za podmínky dožití se časového okamžiku h. Protože délka časového intervalu je pevně dané h a základní hazardní funkce je v čase konstantní a tedy i hazardní funkce je v čase konstantní, pak i tato podmíněná pravděpodobnost je pro jedince s pevně daným regresorem x konstantní. Toto je klíčová skutečnost pro konstrukci simulace, kde budeme simulovat selhání jedinců v čase na základě této pravděpodobnosti.

Mechanizmus simulace

Pro jedince s pevně danou hodnotou regresoru x vygeneruji posloupnost nul a jedniček o délce $\frac{D}{h}$ (počet časových okamžiků) pomocí binomického rozdělení s pravděpodobností $\lambda(x)h$. První jednička v této posloupnosti určuje čas selhání. Pokud se v posloupnosti nevygeneruje žádná jednička, tento jedinec je cenzorován na konci studie. Toto provedu pro všechny jedince, kde každý jedinec je identifikován hodnotou regresoru x. U každého dostávám informaci o době přežití a o selhání či cenzorování. Nyní můžu provést odhad hazardní funkce a odhad Coxova modelu.

Hlavní myšlenka celé simulace je to, jak určím funkční závislost funkce hazardu na regresoru x. Toto lze zapsat vztahem

$$\lambda(x) = \lambda_0 \exp(\beta f(x)).$$

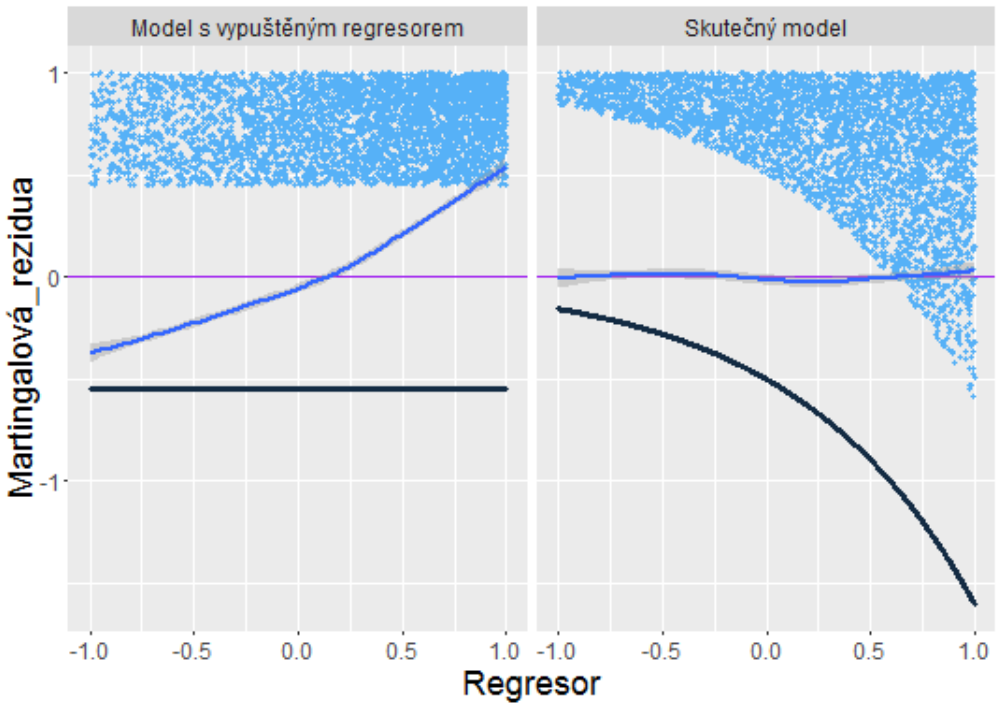
V simulaci budeme postupně uvažovat následující transformace: $f(x) = x$, x^2, x^3, x^4, $\log(x)$, $\exp(x)$, $\sin(x)$, $\cos(x)$

Podmínky simulace

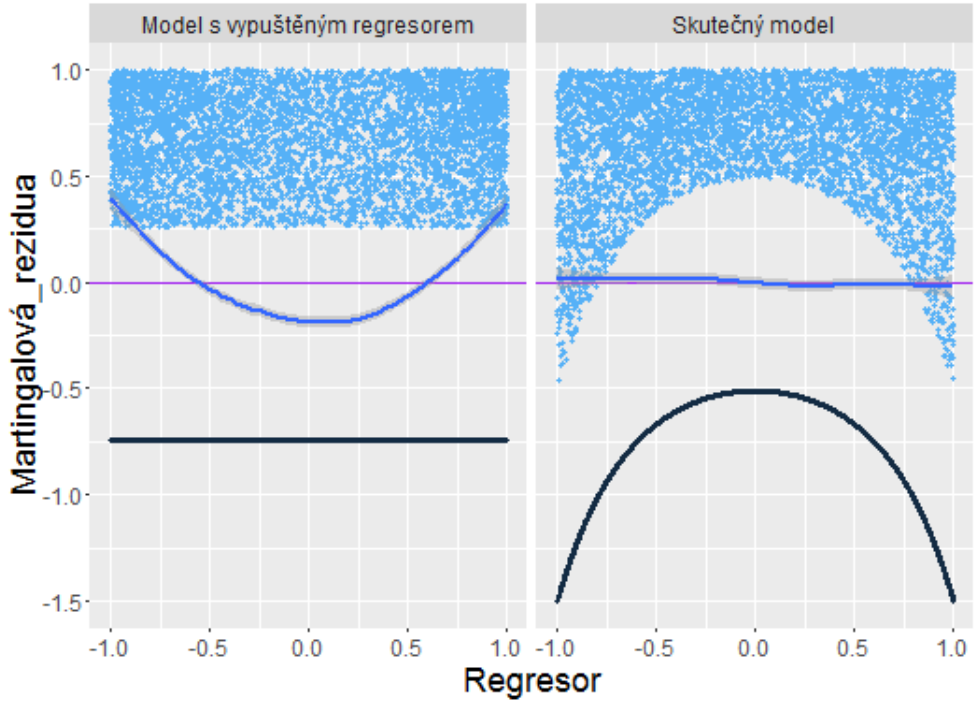
Je ovšem nutné dodržet následující podmínku a tomu i přizpůsobit nastavení parametrů simulace. Je pouze nutné dodržet, aby pravděpodobnost selhání každého jedince byla z intervalu $[0, 1]$.

$$P(T \in [(j-1)\cdot h \quad ; \quad j\cdot h] | T \geq (j-1)\cdot h) = \lambda(x)h = \lambda_0 \exp(\beta f(x))h < 1,$$

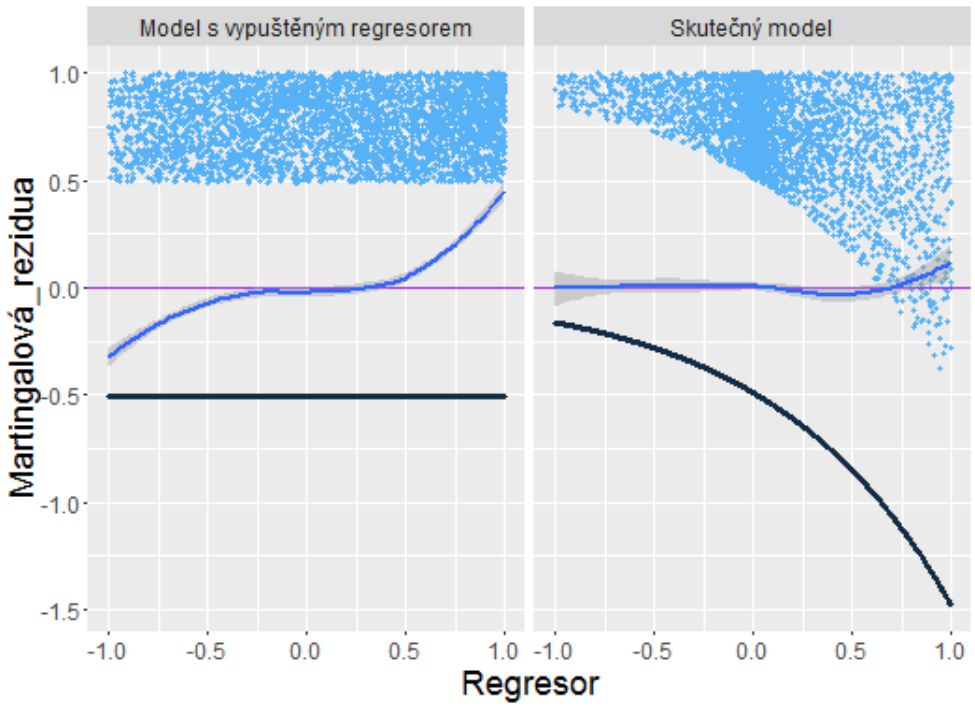
Nyní provedu simulace s různou volbou $f(x)$ a budu sledovat, jak se bude chovat grafická diagnostika martingalových reziduí.



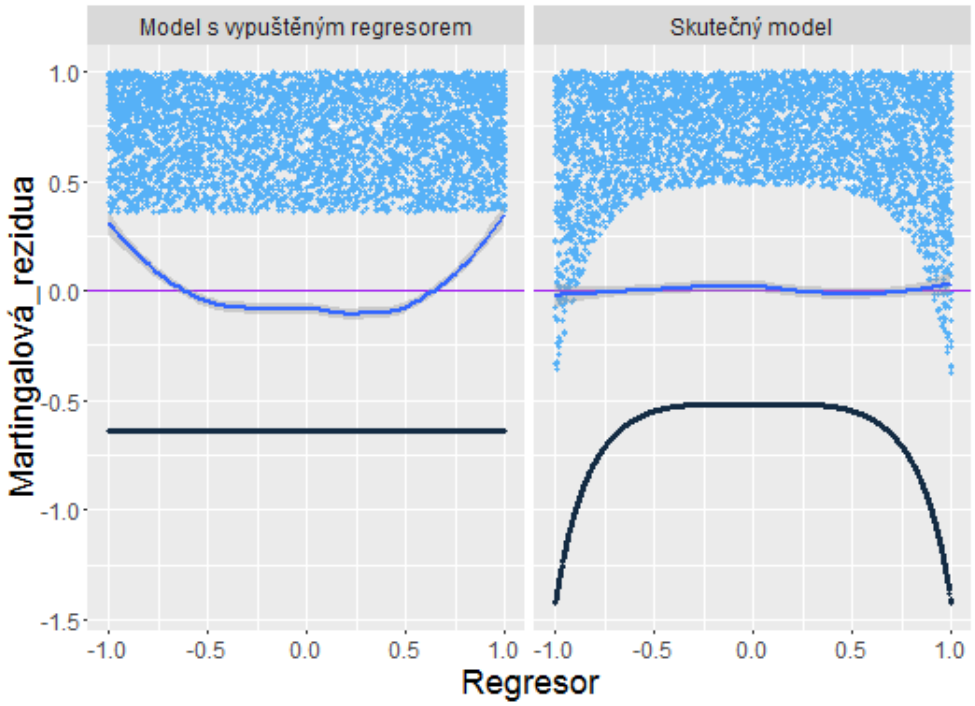
Obrázek 5: Ověření funkčního vlivu proměnné x v případě, že $f(x) = x$



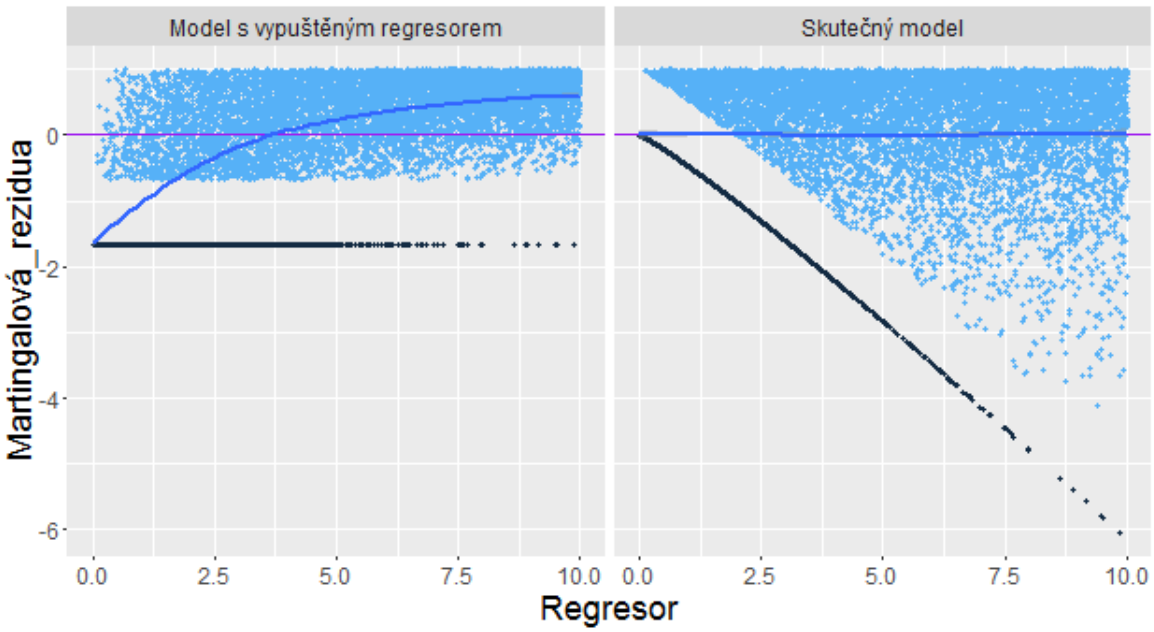
Obrázek 6: Ověření funkčního vlivu proměnné x v případě, že $f(x) = x^2$



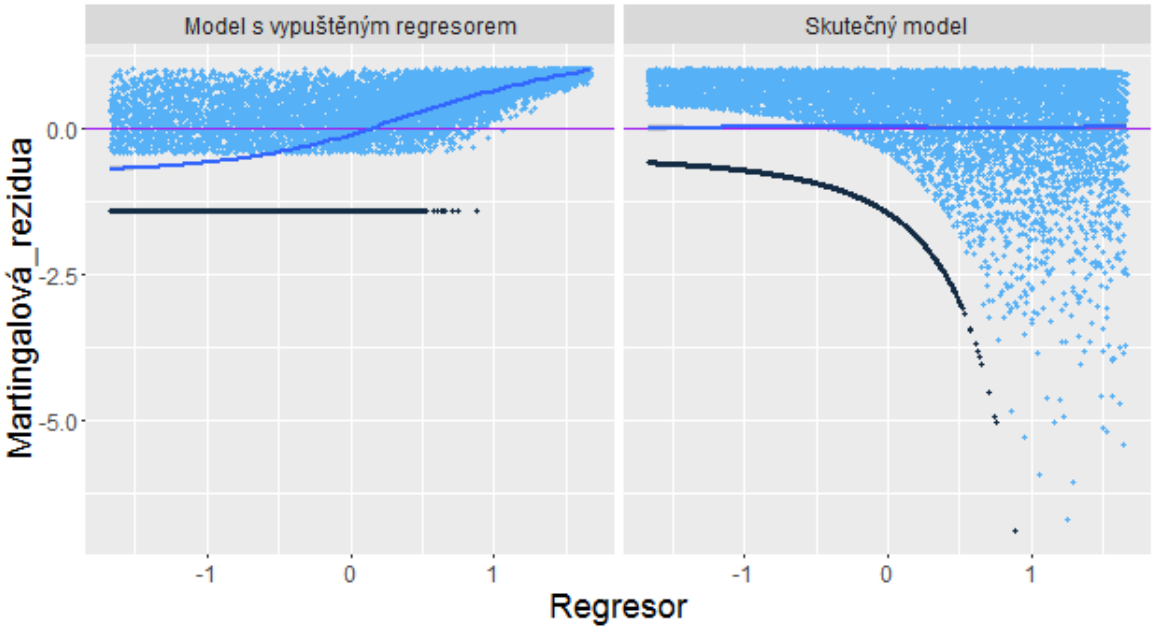
Obrázek 7: Ověření funkčního vlivu proměnné x v případě, že $f(x) = x^3$



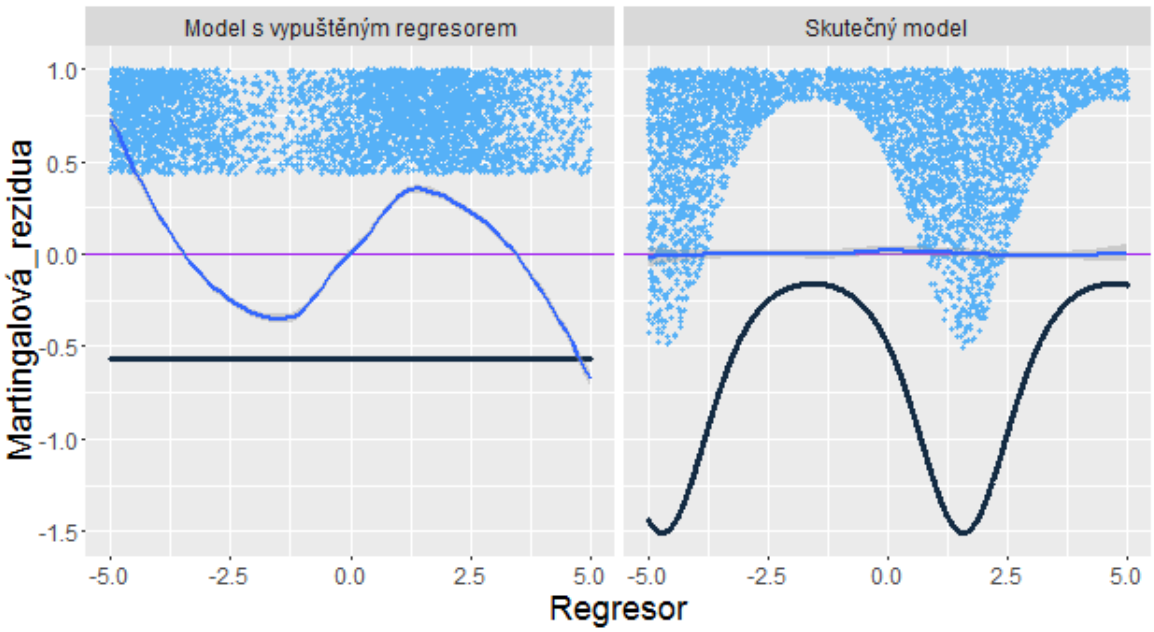
Obrázek 8: Ověření funkčního vlivu proměnné x v případě, že $f(x) = x^4$



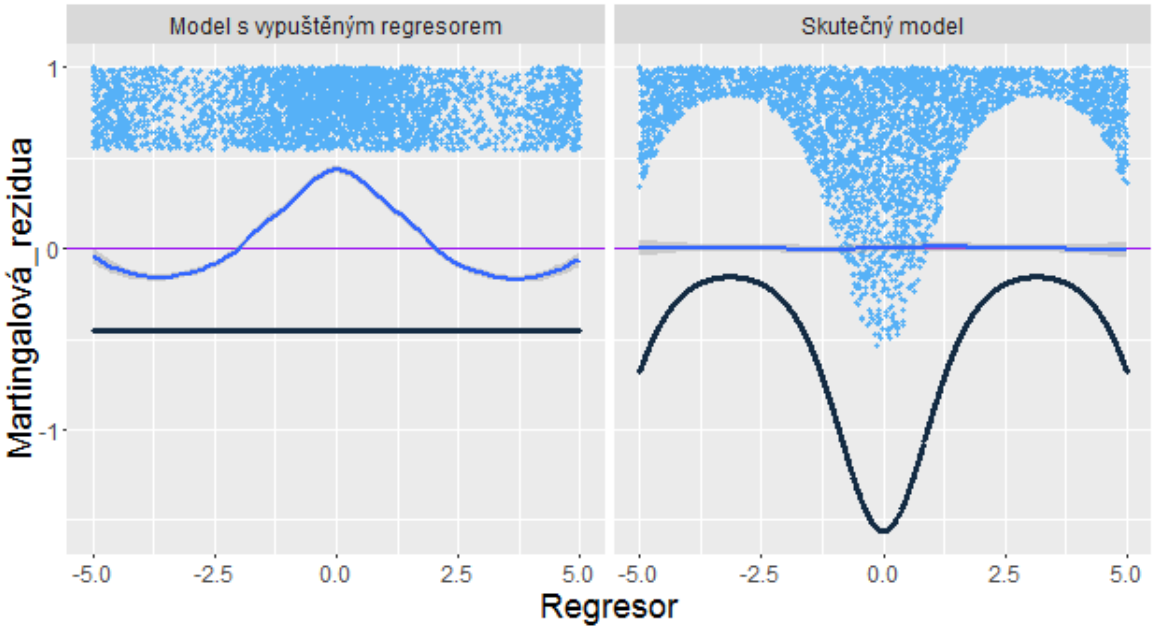
Obrázek 9: Ověření funkčního vlivu proměnné x (v tomto případě bereme hodnoty z intervalu $[0, 10]$) v případě, že $f(x) = \log(x)$



Obrázek 10: Ověření funkčního vlivu proměnné x (v tomto případě bereme hodnoty z intervalu $[-\frac{5}{3}, \frac{5}{3}]$) v případě, že $f(x) = \exp(x)$



Obrázek 11: Ověření funkčního vlivu proměnné x (v tomto případě bereme hodnoty z intervalu [-5,5]) v případě, že $f(x) = \sin(x)$

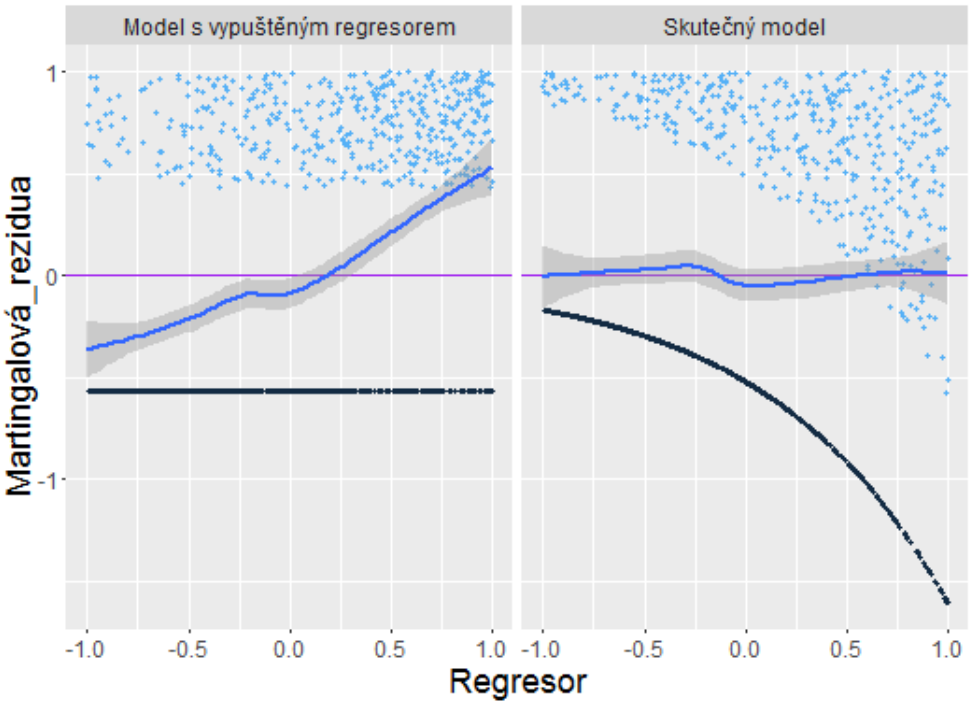


Obrázek 12: Ověření funkčního vlivu proměnné x (v tomto případě bereme hodnoty z intervalu [-5,5]) v případě, že $f(x) = \cos(x)$

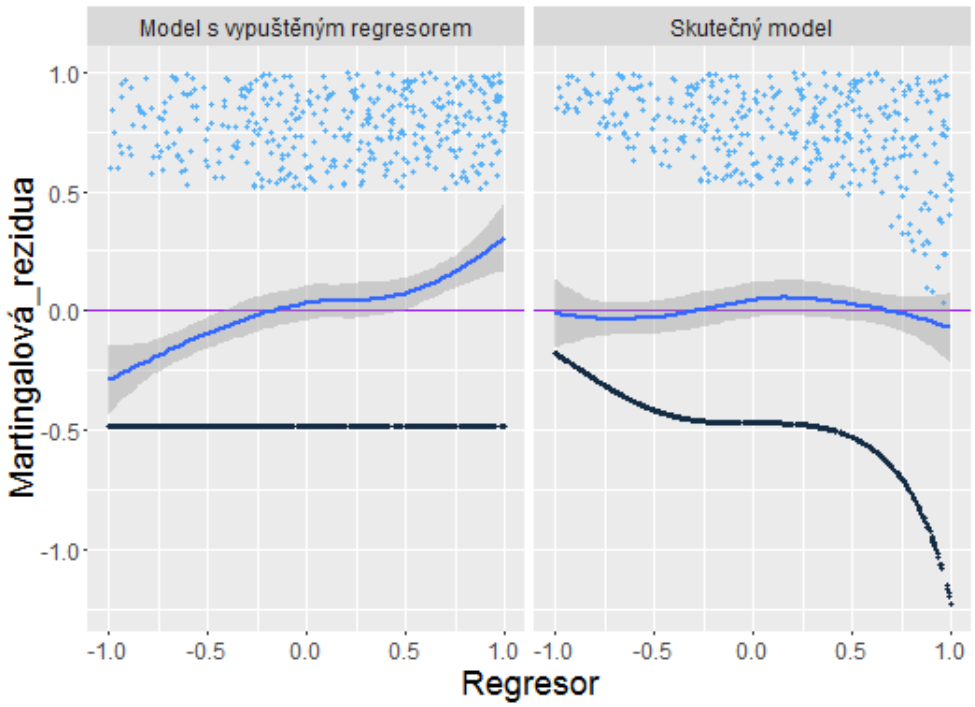
Modré body jsou martingalová rezidua selhavších jedinců, černé cenzorovaných.

Pro lepší přehlednost je u funkcí $\exp, \sin, \cos$ použit jiný nosič regresoru x než u polynomiálních funkcí a to z důvodu podobnosti s lineární funkcí na intervalu $[-1, 1]$. Pro logaritmickou funkci jsem nosič pozměnil z definičních důvodů. Důležité však je, že pro konstrukci grafů a křivky proložení generuji celkem 10000 jedinců. Při tomto počtu už simulace a tvar proložení dává stabilní výsledky. Nasimulovali jsme celkem osm funkčních vztahů a lze říci, že tento funkční vztah se skutečně projevuje na tvaru křivky proložené rezidui.

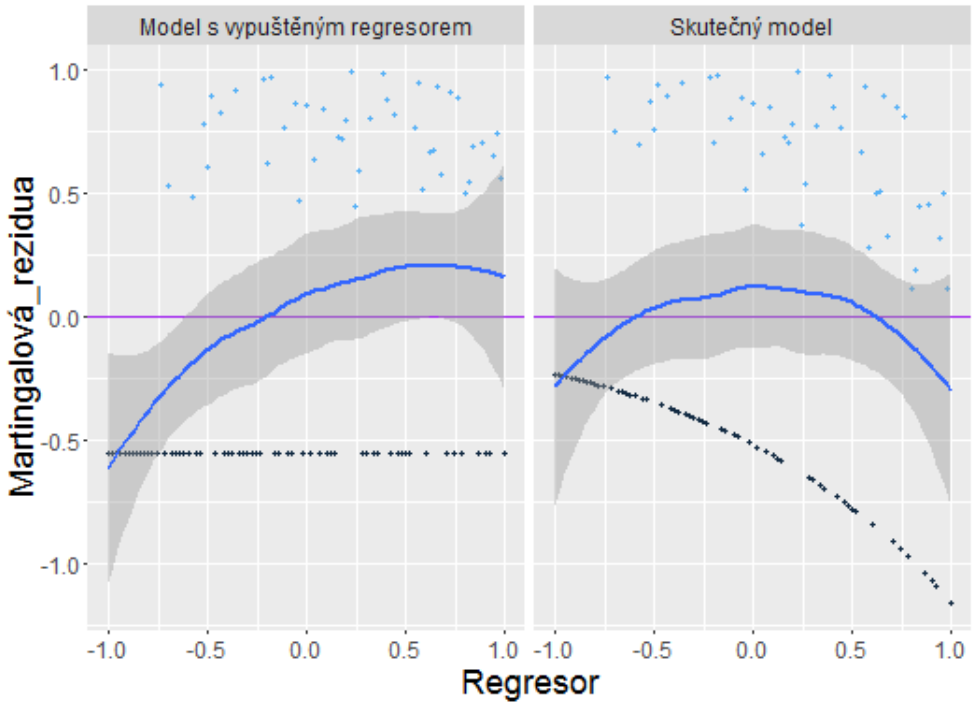
Pro zajímavost vyzkoušíme nasimulovat lineární a kubický vztah pro 1000 a pro 100 jedinců.



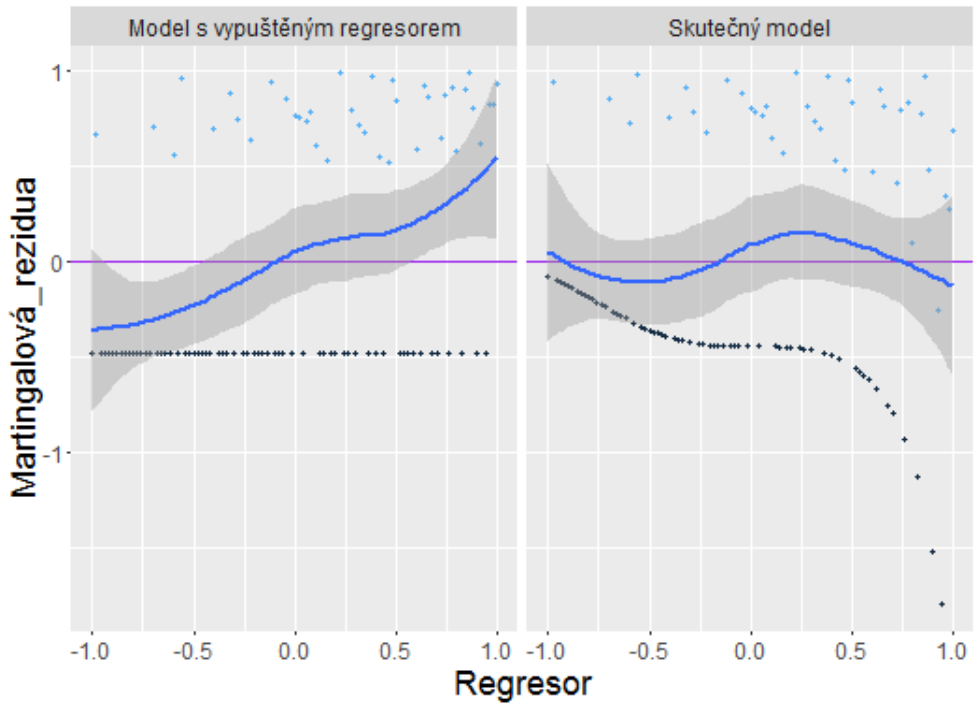
Obrázek 13: Ověření funkčního vlivu proměnné x v případě, že $f(x) = x$ pro 1000 jedinců



Obrázek 14: Ověření funkčního vlivu proměnné x v případě, že $f(x) = x^3$ pro 1000 jedinců



Obrázek 15: Ověření funkčního vlivu proměnné x v případě, že $f(x) = x$ pro 100 jedinců



Obrázek 16: Ověření funkčního vlivu proměnné x v případě, že $f(x) = x^3$ pro 100 jedinců

Můžeme vidět, že pro rozsah výběru 1000 jedinců bychom mohli považovat kubický a lineární funkční vztah za totožný. Pro rozsah výběru 100 jedinců se kubický vztah znovu jeví jako lineární a lineární vztah bychom dokonce mohli považovat za nelineární (např. logaritmus).

3.4. Devianční rezidua

Martingalová rezidua nejsou symetrická kolem nuly ani v případě, kdy odhadnutý model je správný. Tato vlastnost (šikmost) činí metody založené na grafech martingalových těžko interpretovatelné a není úplně snadné něco z těchto grafů vyvozovat. Těžko bychom v naší simulaci odhadovali funkční závislost reziduí na regresoru pouze na základě grafu reziduí bez dodatečného proložení křivkou. Díky jejich vlastnostem nelze martingalová rezidua využít k de-

tekci odlehlých pozorování. Deviance residuals jsou oproti martingalovým reziduím mnohem více symetricky rozdělená kolem nuly. Jejich hlavní diagnostickou vlastností je právě schopnost detekce odlehlých pozorování.

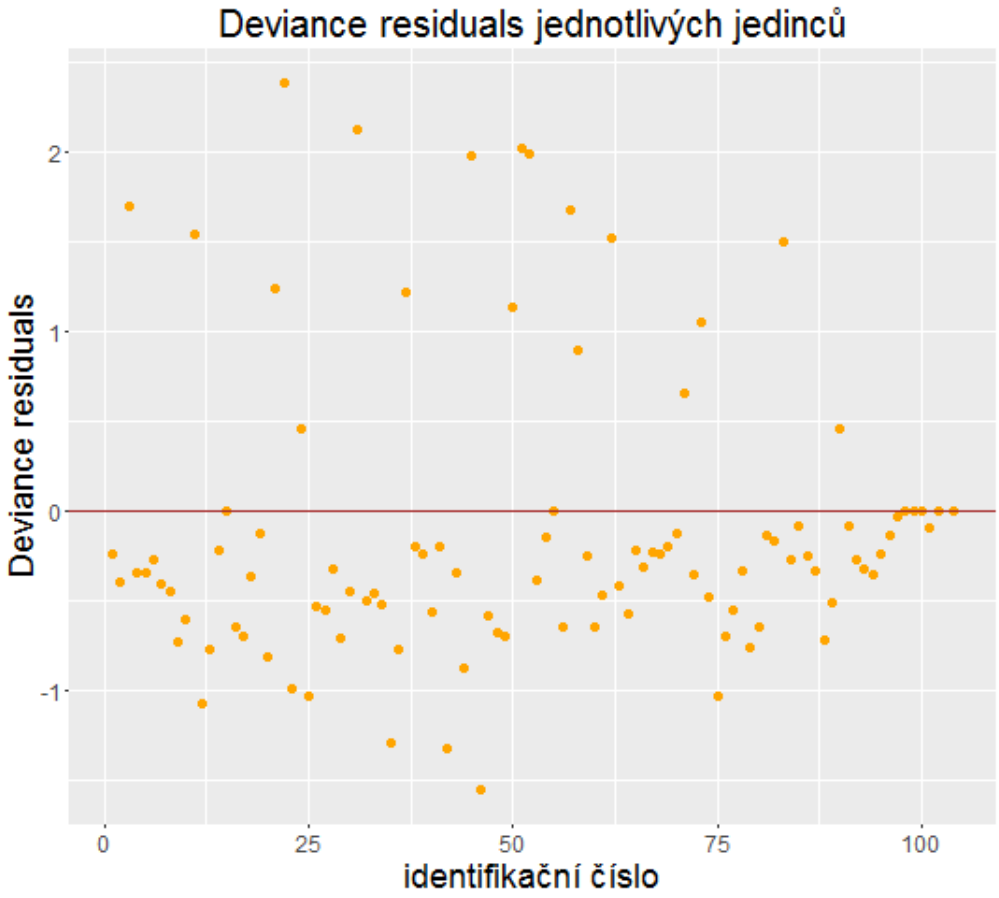
Definice 3.6 (Devianční rezidua) Devianční reziduum je pro i-tého jedince dáno vztahem:

$$d_i = sign(M_i)[-2(M_i + y_i \log(y_i - M_i))]^{1/2}, \tag{3.4.1}$$

kde M_i je martingalové reziduum i-tého jedince a y_i je indikátorem selhání.

Motivace ke konstrukci těchto reziduí je transformovat martingalová rezidua tak, aby jejich hodnoty byly symetrické kolem nuly v případě, že odhadnutý model je vhodný. Již víme, že martingalová rezidua nabývají hodnot z intervalu $[-\infty, 1]$. Pro velké záporné hodnoty M_i bude výraz v hranatých závorkách určen především hodnotou M_i. Druhá odmocnina tohoto výrazu pak jeho hodnotu srazí blíže k nule a d_i bude záporné. Nyní uvažujme rezidua z intervalu $[0, 1]$. Výraz $y_i \log(y_i - M_i)$ bude nenulový pouze pro selhavší subjekty a bude mít hodnotu $\log(1 - M_i)$. Pokud hodnota martingalového rezidua M_i bude blízko 1, hodnota $1 - M_i$ bude blízko nule a $\log(1 - M_i)$ bude nabývat velkých záporných hodnot. Výraz v hranatých závorkách je potom dominován logaritmickou funkcí a hodnoty d_i se pak blíží hodnotě ∞. Druhá odmocnina znova srazí velikost rezidua k nule, kdy teď jeho hodnota bude kladná. Závěrem je nutné dodat, že ačkoliv jsou tato rezidua konstruována ve smyslu symetrie kolem nuly, tak jejich střední hodnota nemusí být nutně rovna nule.

Diagnostické vlastnosti těchto reziduí při detekci odlehlých pozorování jsou obecně považovány za poměrně slabé. Metoda dobře funguje pro detekci těch jedinců, kteří „*žijí moc dlouho*“. V případě jedinců, kteří „*zemřeli moc brzo*“ metoda není spolehlivá.



Obrázek 17: Graf určený pro identifikaci odlehlých pozorování

Na základě grafu lze říci, že námi navržený model modeluje hazardní funkci poměrně dobře. Z grafu není patrné žádné odhadnuté reziduum, které by se oproti zbytku výrazně lišilo.

Závěr

V analýze přežívání se obecně snažíme popsat efekt regresorů na dobu přežití. V každé takovéto studii je nutné přesně definovat počátek studie a příčinu selhání. V této práci uvažujeme pouze data podléhající mechanizmu cenzorování zprava. Pro modelování efektu regresorů na dobu přežití se v analýze přežívání nejčastěji používá Coxův model proporcionálních rizik. Coxův model nemodeluje přímo dobu přežití, ale hazardní funkci, která popisuje rozdělení pravděpodobnosti náhodné veličiny doby přežití. To je podstatný rozdíl proti klasickému regresnímu modelu, který tak činí Coxův model náročnější na porozumění a interpretaci. Coxův model je v literatuře označován jako semiparametrický model. Důvodem je to, že v modelu za neznámé považujeme regresní parametry a základní hazardní funkci. Základní hazardní funkci není nutné nijak specifikovat a ani odhadovat. To, co odhadujeme v Coxově modelu jsou pouze regresní paramety. Je dobré si všimnout, že v základní rovnici Coxova modelu není nijak popsána chybová složka. Ta je už právě zahrnuta v neznámé základní hazardní funkci. Tato semiparametričnost Coxova modelu ovlivňuje také testování hypotéz o parametrech. Při testování nelze vycházet z věrohodnostní funkce, ale je potřeba zvolit přístup založený na parciální věrohodnostní funkci. Pro testování významnosti regresních parametrů odhadnutých pomocí metody maximální parciální věrohodnosti existuje několik asymptotických testů.

Pro ověření předpokladu proporcionality jsme využili dva přístupy. Prvním bylo srovnání Kaplan-Meierova odhadu funkce přežití pro různé kategoriální regresory a druhým Grambschové-Therneaův test a jeho grafická varianta. Pro tento test bylo nutné zavést Schoenfeldova rezidua, jejichž konstrukce je založena přímo na parciální

věrohodnostní funkci. V testu pro daný regresor testujeme hypotézu o závislosti normovaných Schoenfeldových reziduí na dané transformaci času. Ukázali jsme také důležitý vztah, na němž je založena grafická diagnostika Schoenfeldových reziduí. Na základě Grambschové-Therneauova testu se jako problémový regresor jeví Transfúze, což úplně neodpovídá výsledku prvního přístupu, kde se odhady funkcí přežití zdají být paralelní. Jako průkaznější bych považoval výsledek exaktního testu. Ovšem otázka je, jak moc je tento výsledek vypovídající, když v naší studii máme pouze dvacet selhavších jedinců. Při grafickém přístupu bychom porušení proporcionality mohli vyloučit u regresoru Věk, ale u zbylých regresorů bychom mohli mít podezření na porušení tohoto předpokladu.

Pomocí simulace jsme si ukázali, že idea diagnostiky pomocí martingalových reziudí je správná a skutečně lze pozorovat funkční vliv spojitého regresoru na funkci hazardu. Na druhou stranu určité stability výsledků jsme dosáhli až při řádu desetitisíc pozorování. Pokud bychom uvažovali reálnější rozsah výběru v řádu stovek jedinců, tak jsme pomocí simulace ukázali, že metoda není zcela spolehlivá. Jako poslední diagnostický nástroj jsme představili devianční rezidua, která slouží k detekci odlehlých pozorování.

Literatura

[1] Hosmer, D.W.Jr., Lemeshow, S.: *Applied Survival Analysis Regression Modeling of Time to Event Data*. John Wiley & Sons, Inc. 1999

[2] Collett, D.: *Modelling Survival Data in Medical Research*, University of Reading (UK), Springer, 1994

[3] Kalbfleisch, J.D., Prentice R.L.: *The Statistical Analysis of Failure Time Data*. Druhé vydání. John Wiley & Sons, Inc. 2002.

[4] Schoenfeld, D.: *Partial residuals for the proportional hazards regression model*, Biometrika (1982), 69, 1, pp. 239-41.

[5] Kubíčková, V.: *Použití metod analýzy přežití při zkoumání souvislosti chromozomových aberací s výskytem zhoubných novotvarů*, diplomová práce, Přírodovědecká fakulta Univerzity Palackého v Olomouci, 2013

[6] Keele, L.: *Proportionally Difficult: Testing for Nonproportional Hazards in Cox Models*, Department of Political Science, Ohio State University, Columbus, 2010

[7] Marčiny, J.: *Ověřování předpokladů modelu proporcionálního rizika*, Univerzita Karlova v Praze, 2014

[8] Andersen, P.K., Borgan, O., Gill, R.D. and Keiding: *Statistical Models Based on Counting Processes*, Springer-Verlag, New York, 1993