UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Srovnání klasického a bayesovského přístupu k analýze přežívání

Katedra matematické analýzy a aplikací matematiky
Vedoucí diplomové práce: **Mgr. Ondřej Vencálek Ph.D.**
Vypracoval: **Bc. Tereza Brichová**
Studijní program: N1103 Aplikovaná matematika
Studijní obor Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2019

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Tereza Brichová

Název práce: Srovnání klasického a bayesovského přístupu k analýze přežívání

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: Mgr. Ondřej Vencálek Ph.D.

Rok obhajoby práce: 2019

Abstrakt: Cílem diplomové práce je představit a srovnat postupy klasické a bayesovské analýzy přežívání. Oba přístupy poté aplikujeme na reálná data týkající se nádorových onemocnění a porovnáme jejich výsledky.

Klíčová slova: Analýza přežívání, Bayes, Bayesovský přístup

Počet stran: 45

Počet příloh: 0

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Tereza Brichová

Title: Comparison of classical and Bayesian approach to survival analysis

Type of thesis: Master's thesis

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: Mgr. Ondřej Vencálek Ph.D.

The year of presentation: 2019

Abstract: The aim of thesis is to present and compare classical and Bayesian approach to survival analysis. Both approaches will be applied on a real cancer data and then we will compare their results.

Key words: Survival analysis, Bayes, Bayesian approach

Number of pages: 45

Number of appendices: 0

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením pana Mgr. Ondřeje Vencálka Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Seznam obrázků

Seznam tabulek

Poděkování

Ráda bych poděkovala mému vedoucímu diplomové práce Mgr. Ondřeji Vencálkovi, Ph.D. za obětavou spolupráci i čas, který mi věnoval při konzultacích, a také za cenné rady, které pomohly tuto práci dovést do zdárného konce. Poděkování patří i mé rodině a přátelům za podporu během celého studia.

Úvod

Analýza přežívání se zabývá modelováním doby do výskytu nějaké události. Událostí rozumíme například přežití či úmrtí při výskytu nějaké choroby. Analýzu přežívání můžeme také využít ke zkoumání doby než zabere lék. Využití této analýzy můžeme najít i v jiných odvětvích. My však budeme analýzu prezentovat na datech týkajících se rakoviny dutiny ústní a událostí tak pro nás bude skutečně právě úmrtí a tyto dva pojmy tedy budeme v práci volně zaměňovat.

Na úvod si představíme, jak vypadají analyzovaná data a co nás na nich bude zajímat. Poté si předvedeme testy, které při analýze přežívání využijeme. Následovat bude samotná analýza přežívání, kterou nejprve prozkoumáme z pohledu klasické a poté i z pohledu bayesovské statistiky.

Kapitola 1

Data a základní pojmy

1.1 Popis dat

K dispozici máme data týkající se rakoviny v oblasti dutiny ústní, která může být virového a nevirového původu. Zda je rakovina virového či nevirového původu se určuje různými testy, z nichž my máme informace o dvou – RLB a SPF. Tyto testy budeme považovat za hlavní vysvětlující proměnnou. A to z toho důvodu že bychom chtěli vědět, jestli máme léčbu stanovit na základě zjištění, že je rakovina virového původu.

Další důležité faktory, které budeme uvažovat, budou pochopitelně věk pacienta v čase diagnózy, stádium rakoviny, v němž se pacient nachází, velikost tumoru a jeho umístění, zda jde o muže či o ženu a jestli kouří či jestli pije alkohol. Mimo to máme pacienty také rozřazené do rizikových skupin podle odborných článků věnujících se podobnému tématu. Jelikož však nemáme k dispozici přesně tytéž informace o pacientech, byla data rozdělena čtyřmi různými způsoby označenými *Risk Group 1 (g1)* až *Risk Group 4 (g4)*.

Poslední sloupec tabulky udává dobu, po kterou se jedinec ve studii nacházel a prokazatelně žil. Kromě toho máme také informaci o stavu pacienta, tj. jestli, popřípadě z jaké příčiny zemřel, nebo jestli na konci studie stále žil.

1.2 Ukázka dat

sex	lok.	vel.	st.	ter.	stav	kuř.	alk.	RLB	SPF	věk	g1	g2	g3	g4	doba
1	1	4	4	1	5	1	1	16	16	59	3	3	3	3	2.12
1	1	4	4	1	4	1	1	0	0	54	3	3	3	3	0.95
2	1	3	4	2	4	0	0	0	0	70	2	2	2	3	1.52
1	1	2	4	1	5	1	1	0	0	56	2	3	2	3	6.30
1	1	2	3	1	1	NA	NA	0	0	47	NA	NA	NA	NA	13.73

Ještě před samotnou analýzou bude potřeba data trochu upravit. Testy RLB a SPF totiž někdy podávají rozporuplné výsledky, proto můžeme uvažovat analýzu zvlášť pro případ, kdy budou vycházet oba testy pozitivní a zvlášť pro případ, kdy bude aspoň jeden z nich vycházet pozitivní.

1.3 Základní pojmy

- Událost – Existuje několik možností, co můžeme považovat za nastání události. V našem případě budeme pracovat se situací, kdy událostí je úmrtí (tzv. *overall survival*, zkráceně *OS*). Další možností může být například úmrtí z příčiny rakoviny (*disease specific survival*, zkráceně *DSS*). Obě mají svá pro a proti, jelikož OS je ovlivněn úmrtími, která nenastala kvůli studované nemoci, zatímco DSS uvažuje nastání události pouze u jedinců, u kterých došlo prokazatelně k nastání sledované události, kterou je v tomto případě úmrtí na rakovinu. Může nám tak uniknout řada úmrtí, ke kterým dojde z jiné příčiny, která je ovšem důsledkem nemoci.

- Cenzorování – V datech se obvykle vyskytují pozorování, která nevystoupila ze studie kvůli nastání události, ale kvůli tomu, že studie skončila, případně u DSS kvůli tomu, že příčinou úmrtí bylo něco jiného než rakovina. U těchto dat neznáme dobu přežívání. V takovém případě máme dvě možnosti: 1.) můžeme odstranit všechna data, kde nedošlo k úmrtí, avšak jen za cenu obrovské ztráty informace a hlavně úplného zkreslení výsledků, což zřejmě

není právě nejvhodnější, 2.) můžeme provést cenzorování. Podrobnější popis toho, jak se cenzorování provádí, bude vysvětlen dále v textu.

- Funkce přežití – Tato funkce vyjadřuje pravděpodobnost přežití období délky t od začátku sledování (tj. od diagnózy)

$$S(t) = P(T > t).$$

  Funkce přežití je evidentně nerostoucí, zprava spojitá, v bodě $t = 0$ je rovna 1 a limitně pro $t \to \infty$ jde k nule.

- Hazardní funkce (anglicky *hazard function*) – Jedná se o funkci, která vyjadřuje pravděpodobnost úmrtí v daném okamžiku t pro člověka, který se dožil okamžiku jen těsně před tím. Symbolicky ji můžeme vyjádřit v případě, že uvažujeme diskrétní čas, jako

$$h(t) = P(T = t | T \geq t),$$

  a v případě, že uvažujeme spojitý čas, jako

$$h(t) = -\frac{\partial \ln S_t}{\partial t}.$$

Kapitola 2

Použité metody

2.1 Klasický přístup k analýze přežívání

Analýza přežívání představuje v podstatě regresní model, který se liší od ostatních regresních modelů tím, že v sobě zahrnuje proces stárnutí. Jinak řečeno, tato analýza umožňuje v modelu zahrnout také to, že člověk v průběhu času stárne, čímž se mění pravděpodobnost jeho přežití. Chceme-li tedy provádět analýzu přežívání, musíme nejprve zavést *hazardní funkci*, která nám umožňuje tento faktor zahrnout do modelu. Základním modelem je model s exponenciálním rozdělením pravděpodobnosti dob přežití, který se dá lehce odvodit.

Uvažujme lineární regresní model, v němž budeme modelovat závislost logaritmu času na nějakém známém regresoru (např. věk). Pro jednoduchost se zatím budeme zabývat jen jediným regresorem a na situaci s více proměnnými úvahy zobecníme až v pozdější sekci. Uvažovaný model tedy bude popsaný rovnicí

$$y = \beta_0 + \beta_1 x + \sigma \cdot \varepsilon^*,$$

kde $y = \ln(t)$ a $\varepsilon^* = \ln(\varepsilon)$. Vyjádříme-li z této rovnice proměnnou t, dostaneme

$$t = e^{(\beta_0 + \beta_1 x)} \cdot \varepsilon^{\sigma}.$$

Konkrétní tvar hazardní funkce závisí tedy na rozdělení pravděpodobnosti chybové složky. Exponenciální rozdělení časů přežití dostaneme, jestliže zvolíme $\sigma = 1$. Střední hodnotu t potom získáme jako $\frac{1}{\lambda}$ a dostaneme tedy

$$\frac{1}{\lambda} = e^{\beta_0 + \beta_1 x},$$

odkud můžeme vyjádřit $\lambda = e^{-(\beta_0+\beta_1 x)}$, což je zároveň i hazardní funkcí v exponenciálním modelu pro analýzu přežívání, neboť, jak jsme si uvedli výše, funkci přežití získáme jako $1 - F(t)$, kde $F(t) = 1 - e^{-\lambda t}$ je distribuční funkce exponenciálního rozdělení. Platí tedy $S(t) = e^{-\lambda t}$, respektive $\ln S(t) = -\lambda t$. Derivací podle t a vynásobením mínus jedničkou dostaneme, že hazardní funkce je rovna přímo parametru λ z exponenciálního rozdělení časů přežití.

Zde bychom si měli všimnout dvou věcí – uvedená hazardní funkce nezávisí na čase t (tj. nerozlišuje, jestli objekt sledujeme den, týden či celé roky) a je převrácenou hodnotou nenáhodné složky z regresního modelu. První uvedená vlastnost může být v praxi překážkou, jelikož délka doby studie může být významným faktorem. Proto si obvykle nevystačíme jen s exponenciálním rozdělením a model zobecníme na případ, kdy $\sigma \neq 1$ a chybová složka se tak řídí *Weibullovým rozdělením* s parametry α a σ. Hazardní funkce dostane následující podobu

$$h(t, x, \boldsymbol{\beta}, \lambda) = \frac{\lambda t^{\lambda-1}}{(e^{\beta_0+\beta_1 x})^{\lambda}}, \tag{2.1}$$

kde $\lambda = \frac{1}{\sigma}$ kvůli přehlednějšímu zápisu. Budeme-li se na uvedenou funkci dívat jako na funkci proměnné času, dostaneme, že pro $\lambda > 1$ (resp. $\sigma < 1$) hazardní funkce roste s rostoucí dobou přežívání, zatímco pro $\lambda < 1$ ($\sigma > 1$) bude hazardní funkce s přibývajícím časem klesat. Ani tato funkce ovšem není univerzální, neboť předpokládá, že vývoj funkce v závislosti na čase je monotónní. To znamená, že si neporadí například se situací, kdy hazardní funkce do určitého času roste a poté klesá.

Obecně od hazardní funkce můžeme požadovat splnění dvou cílů. Nejprve chceme aby vyjadřovala, jak se mění pravděpodobnost přežití v závislosti na čase, tzn. jaká je pravděpodobnost přežití v každém čase t trvání studie. Druhým cílem je zjistit, jak se tato pravděpodobnost mění v závislosti na měnících se hodnotách regresorů. Často si vystačíme právě jen se splněním druhého cíle, např. pokud chceme znát odpověď na otázku, jaký vliv na přežívání má nová terapie. V tom případě nám může posloužit semiparametrický model, který nemá předpoklady

na rozdělení chybové složky.

Semiparametrické modely předpokládají, že můžeme hazardní funkci vyjádřit ve tvaru

$$h(t, x, \boldsymbol{\beta}) = h_0(t) r(x, \beta_1),$$

kde $h_0(t)$, tzv. *baseline hazard function*, vyjadřuje závislost hazardní funkce na čase (a její součástí je i člen e^{β_0}), zatímco $r(x, \beta_1)$ vyjadřuje závislost této funkce na hodnotě regresoru x. Obě funkce přitom musíme volit tak, aby $h(t, x, \boldsymbol{\beta}) > 0$ pro všechna $t > 0$. Nyní můžeme zavést tzv. *hazard ratio*, které nám bude vyjadřovat poměr mezi dvěma hazardními funkcemi následovně:

$$HR(t, x_1, x_0) = \frac{h(t, x_1, \boldsymbol{\beta})}{h(t, x_0, \boldsymbol{\beta})} = \frac{h_0(t) r(x_1, \beta_1)}{h_0(t) r(x_0, \beta_1)} = \frac{r(x_1, \beta_1)}{r(x_0, \beta_1)}.$$

Tento poměr tedy nezávisí na vývoji hazardní funkce v čase $h_0(t)$, ale pouze na její závislosti na regresoru v podobě funkce $r(x, \beta_1)$.

S uvedeným modelem přišel jako první Cox v roce 1972, kdy uvedl funkci $r(x, \beta_1)$ ve tvaru $e^{x\beta_1}$. Hazardní funkce a hazard ratio v Coxově modelu tedy vypadají následovně:

$$h(t, x, \boldsymbol{\beta}) = h_0(t) \cdot e^{x\beta_1}. \tag{2.2}$$

$$HR(t, x_1, x_0) = e^{\beta_1 (x_1 - x_0)}. \tag{2.3}$$

Takto získané hazard ratio můžeme interpretovat jako poměr relativního rizika. Jestliže tedy hazard ratio pro hazardní funkci žen a mužů vyjde 0.51, znamená to, že riziko úmrtí u mužů je v kterémkoliv okamžiku přibližně dvakrát větší než u žen.

Podívejme se nyní, jak bude vypadat funkce přežívání pro hazardní funkci ve tvaru $h_0(t) \cdot e^{x\beta_1}$. K tomu potřebujeme vědět, že za předpokladu absolutní spojitosti proměnné t můžeme funkci přežívání vyjádřit ve tvaru

$$S(t) = e^{-H(t, x, \boldsymbol{\beta})},$$

kde $H(t, x, \boldsymbol{\beta})$ je kumulativní distribuční funkce v čase t pro pozorování s danými

hodnotami regresorů x. Díky absolutní spojitosti proměnné t pak můžeme vyjádřit

$$H(t,x,\boldsymbol{\beta}) = \int_0^t h(u,x,\boldsymbol{\beta})du = \int_0^t h_0(u)\cdot r(x,\beta_1)du = r(x,\beta_1)\cdot\int_0^t h_0(u)du = \\ = r(x,\beta_1)\cdot H_0(t). \tag{2.4}$$

Kumulativní distribuční funkci tedy můžeme zapsat jako součin základní kumulativní distribuční funkce $H_0(t)$ a funkce $r(x,\beta_1)$ pro subjekt s hodnotami regresorů x. Obecná semiparametrická funkce přežití tedy bude vypadat následovně:

$$S(t,x,\boldsymbol{\beta}) = e^{-r(x,\beta_1)H_0(t)} = [e^{-H_0(t)}]^{r(x,\beta_1)} = [S_0(t)]^{r(x,\beta_1)},$$

kde $S_0(t) = e^{-H_0(t)}$ budeme značit základní funkci přežití. V případě Coxova modelu tak dostaneme:

$$S(t,x,\boldsymbol{\beta}) = [S_0(t)]^{e^{x\beta_1}}.$$

Běžně se regresní modely sestavují tak, aby maximalizovaly funkci věrohodnosti, ale v případě Coxova modelu vypadá příslušná funkce věrohodnosti takto (její odvození je popsáno v [1]):

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n}[h(t_i,x_i,\boldsymbol{\beta})]^{c_i}\cdot[S(t_i,x_i,\boldsymbol{\beta})],$$

respektive

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n}\big(c_i\cdot\ln[h_0(t_i)] + c_i x_i\beta_1 + e^{x_i\beta_1}\ln[S_0(t_i)]\big),$$

kde x_i je hodnota regresoru u i-tého pozorování a c_i vypovídá o tom, zda u i-tého pozorování došlo k události ($c_i = 1$), nebo bylo potřeba cenzorovat ($c_i = 0$). Zásadní problém vyvstává kvůli tomu, že nemáme konkrétní tvar základní hazardní funkce a funkce přežití, proto se místo funkce věrohodnosti používá parciální funkce věrohodnosti (anglicky *partial likelihood function*), která má následující podobu:

$$l_p(\beta_1) = \prod_{i=1}^{n}\left[\frac{e^{x_i\beta_1}}{\sum_{j\in R(t_i)}e^{x_j\beta_1}}\right]^{c_i},$$

kde výraz ve jmenovateli sčítáme přes všechny objekty v rizikové skupině v čase t_i. Rizikovou skupinou $R(t_i)$ přitom rozumíme skupinu pozorování, u níž je doba přežití či doba do cenzorování větší nebo rovna danému časovému okamžiku t_i. Funkce se obvykle uvádí ve tvaru bez výrazů, pro něž $c_i = 0$.

$$l_p(\beta_1) = \prod_{i=1}^{m} \frac{e^{x_{(i)}\beta_1}}{\sum_{j \in R(t_{(i)})} e^{x_j \beta_1}}, \tag{2.5}$$

kde m značí počet různých časů přežití a $x_{(i)}$ je hodnota regresoru pro subjekt s uspořádaným časem přežití $t_{(i)}$. Logaritmus parciální funkce věrohodnosti pak bude vypadat takto:

$$L_p(\beta_1) = \sum_{i=1}^{m} \left(x_{(i)}\beta_1 - \ln \Big[\sum_{j \in R(t_{(i)})} e^{x_j \beta_1} \Big] \right).$$

Zderivováním podle β_1 a srovnáním s nulou získáme následující rovnici (detailní odvození je popsáno v [1]):

$$\sum_{i=1}^{m} \left(x_{(i)} - \bar{x}_{w_i} \right) = 0,$$

kde

$$\bar{x}_{w_i} = \sum_{j \in R(t_{(i)})} w_{ij}(\beta_1) x_j, \qquad\qquad w_{ij}(\beta_1) = \frac{e^{x_j \beta_1}}{\sum_{l \in R(t_i)} e^{x_l \beta_1}},$$

odkud můžeme spočítat odhad parametru β_1 (budeme jej značit $\hat{\beta}_1$).

Ukazuje se, že takto získaný odhad má stejné distribuční vlastnosti jako odhady získané pomocí metody maximální věrohodnosti. Mezi ně patří i to, že odhad asymptotického rozptylu tohoto odhadu můžeme spočítat jako záporně vzatou převrácenou hodnotu druhé derivace logaritmu dílčí funkce věrohodnosti v bodě $\hat{\beta}_1$. S využitím výše uvedeného vztahu pro výpočet $w_{ij}(\beta_1)$ dostaneme

$$\frac{\partial^2 L_p(\beta_1)}{\partial^2 \beta_1} = -\sum_{i=1}^{m} \sum_{j \in R(t_{(i)})} w_{ij}(\beta_1)(x_j - \bar{x}_{w_i})^2.$$

Vynásobíme-li výraz na pravé straně mínus jednou, dostaneme tzv. *pozorovanou informaci* $I(\beta_1) = -\frac{\partial^2 L_p(\beta_1)}{\partial^2 \beta_1}$ (v případě vektoru parametrů $\boldsymbol{\beta}$ hovoříme o *matici pozorované informace*, kterou značíme $\boldsymbol{I(\beta)}$). Odhad asymptotické varianční matice je tedy převrácenou hodnotou pozorované informace v bodě $\beta = \hat{\beta}$.

$$\hat{var}\hat{\boldsymbol{\beta}} = I(\hat{\boldsymbol{\beta}})^{-1}.$$

V praxi obvykle potřebujeme znát směrodatnou odchylku odhadů (kvůli testům významnosti), tu získáme odmocněním rozptylu.

2.1.1 Testy významnosti, konfidenční intervaly

V analýze přežívání se nejčastěji vyskytují tři testy významnosti pro jednotlivé parametry (všechny vyžadují dostatečně velké množství necenzorovaných pozorování):

- Test založený na poměru věrohodností (partial-likelihood ratio test)

  ▷ Testová statistika má tvar: $G = 2 \cdot [L_p(\hat{\beta}_1) - L_p(0)]$
  $L_p(0) = -\sum_{i=1}^{m} \ln(n_i)$ ... hodnota maxima logaritmu parciální funkce věrohodnosti za podmínky $\beta_1 = 0$
  $L_p(\hat{\beta}_1)$ ... maximum logaritmu parciální funkce věrohodnosti v modelu s odhadem regresního parametru $\hat{\beta}_1$
  Za platnosti nulové hypotézy platí $G \dot{\sim} \chi^2_1$

- Waldův test

  ▷ Testová statistika má tvar: $z = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)}$
  Za platnosti nulové hypotézy platí $z \dot{\sim} N(0,1)$

- Skórový test

  ▷ Testová statistika má tvar: $z^* = \frac{\partial L_p / \partial \beta_1}{\sqrt{I(\beta_1)}}|_{\beta_1=0}$
  Za platnosti nulové hypotézy platí $z^* \dot{\sim} N(0,1)$

V praxi vycházejí obvykle výsledky všech tří testů velmi podobně, pokud by však došlo k rozporu mezi nimi, doporučuje se rozhodnout na základě *partial-likelihood ratio testu*. Waldův test se využívá spíše u modelů s více vysvětlujícími proměnnými, kde slouží jako kritérium pro výběr významných regresorů.

Uvedené testy můžeme využít také při konstrukci konfidenčních intervalů parametrů β_1. Jako příklad si můžeme uvést využití Waldovy statistiky, kdy víme, že odhady mají asymptoticky normální rozdělení pravděpodobnosti se směrodatnou odchylkou $\sqrt{I(\hat{\beta}_1)^{-1}}$. Koncové body $100(1-\alpha)\%$ intervalu v takovém případě dostaneme jako:

$$\hat{\beta}_1 \pm u_{1-\frac{\alpha}{2}} \hat{SE}(\hat{\beta}_1),$$

kde $u_{1-\frac{\alpha}{2}}$ značí $100(1-\frac{\alpha}{2})\%$ kvantil normovaného normálního rozdělení.

2.1.2 Model s více vysvětlujícími proměnnými

Uvažujme nyní případ s p regresory, jejichž hodnoty jsou změřeny na každém pozorování a po celou dobu studie se nemění. Pro každé pozorování nyní budeme mít k dispozici uspořádanou trojici hodnot $(t_i, \boldsymbol{x}_i, c_i)$, kde t_i je pozorovaný čas přežívání, $\boldsymbol{x}_i' = (x_{i1}, \ldots, x_{ip})$ je vektor regresorů a c_i nese informaci o tom, zda bylo pozorování cenzorované, nebo ne. Parciální funkci věrohodnosti poté dostaneme, když ve vztahu 2.5 nahradíme jediný regresor $x_{(i)}$ vektorem regresorů $\boldsymbol{x}_{(i)}$ a koeficient β vektorem koeficientů $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$.

Podobně jako v případě modelu s jednou vysvětlující proměnnou, dostaneme maximálně věrohodné odhady vektoru koeficientů $\boldsymbol{\beta}$ tak, že výraz zderivujeme podle každé složky tohoto vektoru a výsledné výrazy položíme rovny nule. Rovnice získaná derivací podle k-té složky bude vypadat takto:

$$\frac{\partial L_p(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{i=1}^{m} \left[x_{(ik)} - \frac{\sum_{j \in R(t_{(i)})} x_{jk} e^{\boldsymbol{x}_j' \boldsymbol{\beta}}}{\sum_{j \in R(t_{(i)})} e^{\boldsymbol{x}_j' \boldsymbol{\beta}}} \right] = \sum_{i=1}^{m} \left[x_{(ik)} - \bar{x}_{w_i k} \right],$$

kde

$$w_{ij}(\boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}_j' \boldsymbol{\beta}}}{\sum_{l \in R(t_{(i)})} e^{\boldsymbol{x}_l' \boldsymbol{\beta}}} \qquad \bar{x}_{w_i k} = \sum_{j \in R(t_{(i)})} w_{ij}(\boldsymbol{\beta}) x_{jk}$$

a $x_{(ik)}$ označuje hodnotu k-tého regresoru pro subjekt s uspořádanou dobou přežití $t_{(i)}$. Takto získaný odhad koeficientů $\boldsymbol{\beta}$ budeme značit $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$. Vícerozměrnou analogií pozorované informace bude informační matice, kterou získáme jako

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

a stejně jako v jednorozměrném případě budeme s její pomocí odhadovat variační matici parametrů $\boldsymbol{\beta}$ jako

$$\hat{var}(\hat{\boldsymbol{\beta}}) = \boldsymbol{I}(\hat{\boldsymbol{\beta}})^{-1}.$$

Jednoduché zobecnění pro případ s více regresory lze provést rovněž u testů významnosti vektoru koeficientů. Nulová hypotéza pro model s p vysvětlujícími proměnnými bude mít podobu $\beta_1 = \cdots = \beta_p = 0$. V případě *partial-likelihood testu* bude mít testová statistika opět tvar

$$G = 2 \cdot [L_p(\hat{\boldsymbol{\beta}}) - L_p(\mathbf{0})]$$

a za platnosti nulové hypotézy se bude asymptoticky řídit rozdělením χ^2_p. Zamítnutí nulové hypotézy indikuje, že alespoň jeden ze zkoumaných regresorů statisticky významně souvisí s dobou přežívání. Analogie zbylých dvou testů se dají získat s pomocí maticových výpočtů, které jsou však v praxi nepohodlné a spokojíme se proto s uvedeným *partial-likelihood ratio testem*. Ten se využívá také k porovnání dvou modelů, podobně jako se v lineární regresi používá F-test. Výsledná statistika má za nulové hypotézy (platí podmodel) χ^2 rozdělení s počtem stupňů volnosti daným počtem parametrů, o které se modely liší. Je však potřeba, aby byly oba modely odhadnuty na základě stejných pozorování.

Výše uvedené metody založené na parciální funkci věrohodnosti vycházejí z předpokladu, že v pozorovaných dobách přežití nedošlo ke shodám. To se v praxi stává spíš zřídka a zavedly se proto některé modifikace. Jedna z možností je předpokládat, že ke shodám dochází kvůli nepřesnostem v měření, a spočítat přesný partial likelihood tak, že zahrneme všechna možná uspořádání těchto shod.

Jinou cestou je využít Breslowovu (2.6) nebo Efronovu aproximaci (2.7):

$$l_{p1}(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{e^{x'_{(i)+}\boldsymbol{\beta}}}{\left[\sum_{j\in R(t_{(i)}} e^{x'_j\boldsymbol{\beta}}\right]^{d_i}} \tag{2.6}$$

$$l_{p2}(\beta) = \prod_{i=1}^{m} \frac{e^{x'_{(i)+}\boldsymbol{\beta}}}{\prod_{k=1}^{d_i}\left[\sum_{j\in R(t_{(i)}} e^{x'_j\boldsymbol{\beta}} - \frac{k-1}{d_i}\sum_{j\in D(t_{(i)}} e^{x'_j\boldsymbol{\beta}}\right]}, \tag{2.7}$$

kde d_i značí počet pozorování se stejnou dobou přežití $t_{(i)}$ a $x_{(i)+} = \sum_{j\in D(t_{(i)})} x_j$, kde $D(t_{(i)})$ je skupina pozorování s dobou přežití $t_{(i)}$. Jinými slovy, $x_{(i)+}$ je rovno součtu hodnot regresorů u pozorování s dobou přežití $t_{(i)}$. S uvedenými aproximacemi bychom dále pracovali stejně jako s parciální věrohodnostní funkcí bez shod.

Efronova varianta dává nepatrně lepší aproximaci přesné parciální věrohodnostní funkce, její výpočet je však složitější a mnohé statistické softwary proto využívají k výpočtům Breslowovu variantu. Pro data bez shod dají obě aproximace stejnou hodnotu jako parciální věrohodnostní funkce, která shody neuvažuje.

2.1.3 Coxova funkce přežití

Připomeneme-li si Coxovu funkci přežití

$$S(t, \boldsymbol{x}, \boldsymbol{\beta}) = [S_0(t)]^{e^{\boldsymbol{x}'\boldsymbol{\beta}}},$$

můžeme vidět, že po odhadu regresních koeficientů už nám ke stanovení této funkce zbývá jen odhadnout základní funkci přežití $S_0(t)$. K tomu nás dovedou následující úvahy:

Podmíněnou pravděpodobnost přežití do času $t_{(i)}$ osoby, která žila v čase $t_{(i-1)}$ můžeme zřejmě spočítat jako $\frac{S(t_{(i)})}{S(t_{(i-1)})}$. Definujme podmíněnou základní pravděpodobnost přežití jako $\alpha_i = \frac{S_0(t_{(i)})}{S_0(t_{(i-1)})}$. Potom

$$\frac{S(t_{(i)}, \boldsymbol{x}, \boldsymbol{\beta})}{S(t_{(i-1)}, \boldsymbol{x}, \boldsymbol{\beta})} = \left[\frac{[S_0(t_{(i)})]^{e^{\boldsymbol{x}'\boldsymbol{\beta}}}}{[S_0(t_{(i-1)})]^{e^{\boldsymbol{x}'\boldsymbol{\beta}}}}\right] = \left[\frac{S_0(t_{(i)})}{S_0(t_{(i-1)})}\right]^{e^{\boldsymbol{x}'\boldsymbol{\beta}}} = \alpha_i^{e^{\boldsymbol{x}'\boldsymbol{\beta}}} \tag{2.8}$$

Při počítání maximálně věrohodného odhadu uvažujeme jako $\boldsymbol{\beta}$ dříve uvedený odhad získaný pomocí parciální funkce věrohodnosti, který jsme si označili $\hat{\boldsymbol{\beta}}$. Pro zjednodušení výsledných rovnic (jejich detailnější odvození je popsáno v [5]) zavedeme značení $\hat{\theta}_l = e^{\boldsymbol{x}_l'\hat{\boldsymbol{\beta}}}$. Rovnice pro maximálně věrohodné odhady $\hat{\alpha}_i$ budou vypadat takto:

$$\sum_{l \in D_i} \frac{\hat{\theta}_l}{1 - \alpha_i^{\hat{\theta}_l}} = \sum_{l \in R_i} \hat{\theta}_l \qquad i = 1, \ldots, m, \tag{2.9}$$

kde m je počet různých časů přežití ve studii, R_i značí jedince v rizikové skupině v uspořádané době přežití $t_{(i)}$ a D_i subjekty v rizikové skupině, kteří se dožili právě času $t_{(i)}$.

V případě, že se v datech nevyskytují shody (tj. D_i jsou jednoprvkové množiny), dostaneme

$$\hat{\alpha}_i = \left[1 - \frac{\hat{\theta}_i}{\sum_{l \in R_i} \hat{\theta}_l}\right]^{\hat{\theta}_i^{-1}}. \tag{2.10}$$

Pokud se v datech shody vyskytují, musíme řešení získat pomocí iterativních metod.

Odhad základní funkce přežití dostaneme jako součin jednotlivých odhadů podmíněných zákadních pravděpodobností přežití $\hat{\alpha}_i$,

$$\hat{S}_0(t) = \prod_{i:t_{(i)} \leq t} \hat{\alpha}_i.$$

Jinou cestou je využití Breslowovy aproximace a nahrazení $\hat{\alpha}_i^{\hat{\theta}_l} \approx 1 + \hat{\theta}_l \cdot \ln(\alpha_i)$ na levé straně rovnice 2.9. Řešením bude:

$$\bar{\alpha}_i = \exp\left[\frac{-d_i}{\sum_{l \in R_i} \hat{\theta}_l}\right] \tag{2.11}$$

Základní funkci přežití bychom opět získali jako součin jednotlivých $\bar{\alpha}_i$. Odhad funkce $S(t, \boldsymbol{x}, \boldsymbol{\beta})$ potom dostaneme dosazením odhadu základní funkce přežití a odhadů regresních koeficientů $\hat{\boldsymbol{\beta}}$.

V některých softwarech se můžeme setkat také s odhadem základní hazardní funkce $\hat{h}_0(t_{(i)}) = 1 - \hat{\alpha}_i$. Tyto bodové odhady nejsou příliš přesné a jsou nestabilní,

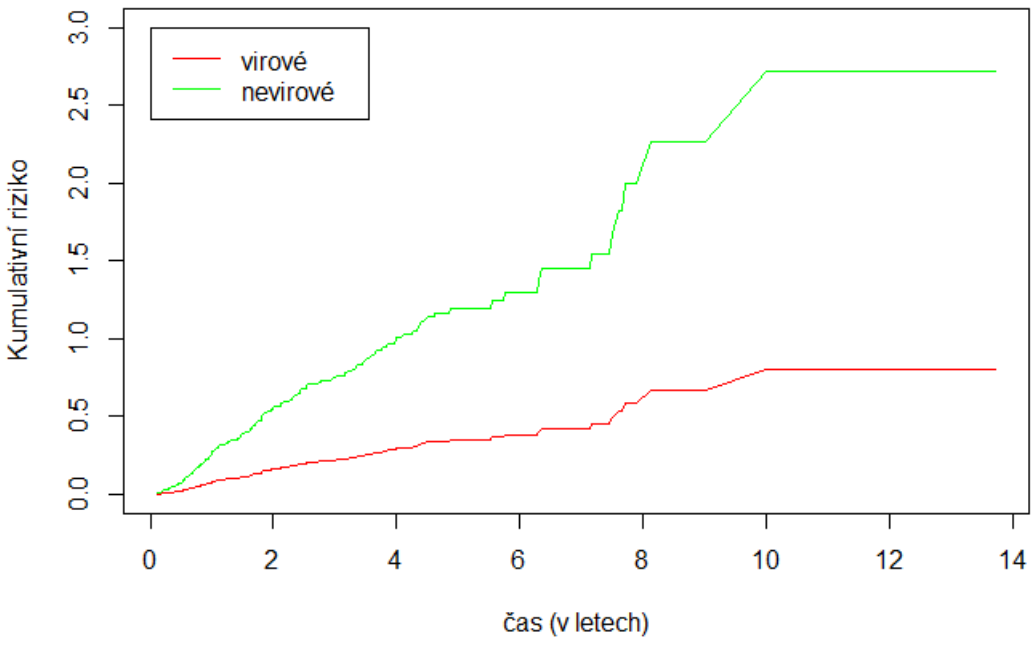
takže se samy o sobě nedají použít. Vhodnější je použít kumulativní základní hazardní funkci, která se dá spočítat ze základní funkce přežívání jako

$$\hat{H}_0(t) = -\ln\left[\hat{S}_0(t)\right].$$

Odhad kumulativní hazardní funkce při daných hodnotách regresorů se spočítá jako

$$\hat{H}(t, \boldsymbol{x}, \hat{\boldsymbol{\beta}}) = -\ln\left[\hat{S}(t, \boldsymbol{x}, \hat{\boldsymbol{\beta}})\right] = -e^{\boldsymbol{x}'\hat{\boldsymbol{\beta}}} \cdot \ln\left[\hat{S}_0(t)\right].$$

Tuto funkci můžeme vykreslit do grafu v závislosti na čase a znázornit tak zkušenost s přežíváním pro jednotlivé hodnoty regresorů, například pro skupinu pacientů s virovým a nevirovým onemocněním, což je vidět na obrázku 2.1.



Obrázek 2.1: Srovnání kumulativní hazardní funkce pro pacienty s virovým nádorovým onemocněním a s nevirovým nádorovým onemocněním

2.1.4 Interpretace proporcionálního hazardního modelu

K interpretaci jednotlivých koeficientů musíme nejprve určit funkční závislost mezi hazardní funkcí a sledovanou závisle proměnnou (pravděpodobností přežití). Z toho totiž vyplyne, jakou transformaci sledované proměnné dostaneme jako lineární kombinaci regresorů. V případě Coxova modelu s jedinou vysvětlující proměnnou (pro více regresorů by výpočet vypadal analogicky) má hazardní funkce tvar

$$h(t, x, \beta) = h_0(t)e^{x\beta}$$

a funkční závislost, tzv. linkovou funkci, tedy získáme jako přirozený logaritmus

$$g(t, x, \beta) = \ln\big[h(t, x, \beta)\big] = \ln\big[h_0(t)\big] + x\beta.$$

Rozdíl linkových funkcí při změně z $x = a$ na $x = b$ bude

$$g(t, x = a, \beta) - g(t, x = b, \beta) = \ln\big[h_0(t)\big] + a\beta - \ln\big[h_0(t)\big] + b\beta = (a - b)\beta.$$

Když výrazy v této rovnici dosadíme do funkce e^x, abychom se zbavili logaritmu, dostaneme

$$HR(t, a, b, \beta) = e^{(a-b)\beta}, \tag{2.12}$$

tedy vztah pro výpočet poměru rizik, se kterým jsme se již seznámili dříve a který hraje významnou roli právě při interpretaci parametrů. Jedná se v podstatě o analogii poměru šancí v logistické regresi. Narozdíl od něj však udává poměr rizika pro subjekty se dvěma různými hodnotami regresoru v každém časovém okamžiku studie a nerozlišuje pouze úmrtí před koncem studie a po něm.

Výběrové rozdělení odhadu poměru rizik je sešikmené doprava, zatímco o samotném parametru β to tolik neplatí a ten má tedy mnohem blíže k normálnímu rozdělení. Proto při konstrukci intervalů spolehlivosti vycházíme z jeho aproximace normálním rozdělením a chceme-li znát interval spolehlivosti pro odhad HR, dosadíme koncové body intervalu pro β do funkce e^x.

$$P\big(\hat{\beta} - \sqrt{\hat{var}(\hat{\beta})}u(0.975) \leq \beta \leq \hat{\beta} + \sqrt{\hat{var}(\hat{\beta})}u(0.975)\big) = 0.95$$

$$P\Big(e^{\hat{\beta}-\sqrt{v\hat{a}r(\hat{\beta})}u(0.975)} \leq HR \leq e^{\hat{\beta}+\sqrt{v\hat{a}r(\hat{\beta})}u(0.975)}\Big) = 0.95,$$

kde $u(0.975)$ značí 97.5% kvantil normovaného normálního rozdělení.

Odhad hazard ratio představuje další možnost, jak zkoumat, zda je koeficient β statisticky významný a to tak, že jej považujeme za významný, jestliže konfidenční interval poměru rizik neobsahuje jedničku.

2.2 Bayesovský přístup k analýze přežívání

U bayesovské analýzy přežívání budeme pracovat s Weibullovým modelem, který předpokládá, že se časy přežívání řídí Weibullovým rozdělením s hustotou

$$f(y|\alpha,\lambda) = \begin{cases} \frac{\alpha}{\lambda}\Big(\frac{y}{\lambda}\Big)^{\alpha-1} e^{-(\frac{y}{\lambda})^{\alpha}} & \text{pro} \quad y > 0 \\ 0 & \text{jinak} \end{cases} \tag{2.13}$$

kde $\alpha > 0$ se nazývá parametr tvaru a $\lambda > 0$ parametr škály. U něj budeme v případě analýzy přežívání uvažovat, že jej lze vyjádřit ve tvaru[2]:

$$\lambda(x_i) = \exp\Big\{-\frac{\mu + \boldsymbol{x}_i'\boldsymbol{\beta}}{\alpha}\Big\}, \tag{2.14}$$

přičemž $\boldsymbol{x}_i$ značí i-tý řádek matice regresorů $\boldsymbol{X}$ uvažovaný jako sloupcový vektor. Z uvedeného předpisu můžeme vidět, že parametr $\lambda(x)$ se bude zvětšovat s rostoucími hodnotami x_{ij}, jestliže bude parametr β_j záporný, a zmenšovat pro β_j kladný. Zároveň platí[3], že střední hodnota Weibullova rozdělení se spočítá jako

$$E(y) = \lambda\Gamma\Big(1+\frac{1}{\alpha}\Big),$$

z čehož můžeme vidět, že průměrná doba přežití roste s rostoucími hodnotami $\lambda(x)$, kde závislost na x má tvar 2.14 (tedy při kladných hodnotách parametrů β_j doba přežití klesá s rostoucími hodnotami regresorů, zatímco při záporných roste).

Zatímco v klasické analýze přežívání bychom hledali parametry $\alpha, \boldsymbol{\beta}$ a μ tak, že bychom maximalizovali věrohodnostní funkci, tj. funkci

$$L(\boldsymbol{y}, \boldsymbol{X}|\alpha,\boldsymbol{\beta},\mu) = \prod_i^n \frac{\alpha}{\lambda_i}\Big(\frac{y_i}{\lambda_i}\Big)^{\alpha-1} e^{-(\frac{y_i}{\lambda_i})^{\alpha}}, \qquad \lambda_i = \exp\Big\{-\frac{\mu+\boldsymbol{x}_i'\boldsymbol{\beta}}{\alpha}\Big\},$$

v bayesovské analýze přežívání budeme pracovat s celou věrohodnostní funkcí, nejen s jejím maximem. Princip bayesovské analýzy spočívá v tom, že stanovíme nějaké své předběžné přesvědčení o hodnotách hledaných parametrů, a to poté vynásobíme funkcí věrohodnosti. Výsledek musíme ještě znormovat, abychom skutečně dostali rozdělení pravděpodobnosti daných parametrů, tj. aby integrál z výsledné hustoty byl roven jedné.

Bayesova věta

Abychom odvodili Bayesovu větu, musíme vyjít ze vzorců pro podmíněnou pravděpodobnost, která říká, že pravděpodobnost náhodného jevu B za dané podmínky, že nastal jev A, lze spočítat jako

$$P(B|A) = \frac{P(B \cap A)}{P(A)}.$$

Analogicky bychom pravděpodobnost $P(A|B)$ získali záměnou symbolů A a B, a z této rovnice bychom mohli vyjádřit pravděpodobnost průniku obou jevů jako

$$P(A \cap B) = P(A|B)P(B),$$

což poté dosadíme do uvedeného vzorce pro výpočet $P(B|A)$ a dostaneme

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}.$$

Pokud máme jevů B_k více a tvoří rozklad jevu A, pak dosadíme do jmenovatele z věty o úplné pravděpodobnosti

$$P(A) = \sum_k P(A|B_k)P(B_k)$$

a dostaneme Bayesovu větu ve tvaru

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_k P(A|B_k)P(B_k)}.$$

My však budeme potřebovat její podobu pro případ, že náhodné jevy A a B_k popisujeme pomocí náhodné veličiny se spojitým rozdělením pravděpodobnosti,

a proto musíme místo pravděpodobnosti v uvedeném vzorci psát hustotu náhodné veličiny:

$$f(B_j|A) = \frac{f(A|B_j)f(B_j)}{\int_k f(A|B_k)f(B_k)dA}.$$

Aplikace Bayesovy věty

Jelikož chceme znát rozdělení pravděpodobnosti parametrů při daných hodnotách pozorování, bude pro nás B_j představovat jev *parametr* α *nabývá hodnoty* j_1, $\beta_l = j_{2l}$ *pro* $l = 1, \ldots, p$, *a* $\mu = j_3$, kde p je délka vektoru $\boldsymbol{\beta}$, zatímco jev A je, že jsme naměřili pozorované hodnoty $\boldsymbol{y}$ při daných hodnotách $\boldsymbol{X}$. Dole potom místo jednoduchého integrálu dostaneme integrál $(p+2)$-rozměrný (máme p parametrů β_l a parametry μ a α). Výraz ve jmenovateli je však konstantní a v praxi se tedy obvykle uvažuje Bayesova věta ve spojitém případě jako

$$f(B_j|A) \propto f(A|B_j)f(B_j), \tag{2.15}$$

přičemž $f(B_j)$ popisuje naše apriorní přesvědčení o hodnotách hledaných parametrů (obvykle uvažujeme hustotu rozdělení pravděpodobnosti s velkým rozptylem, abychom vyjádřili svou nejistotu ohledně hodnoty hledaných parametrů) a $f(A|B_j)$ je funkce věrohodnosti, kterou jsme si uvedli výše. Vztah 2.15 tak můžeme vyjádřit následovně:

$$f_{posterior}(\alpha, \boldsymbol{\beta}, \mu) = kL(\alpha, \boldsymbol{\beta}, \mu|\boldsymbol{y}, \boldsymbol{X})f_{prior}(\alpha, \boldsymbol{\beta}, \mu),$$

přičemž k je normující konstanta, kterou můžeme určit z rovnice

$$\int_K kL(\alpha, \boldsymbol{\beta}, \mu|\boldsymbol{y}, \boldsymbol{X})f_{prior}(\alpha, \boldsymbol{\beta}, \mu)d\alpha d\beta_1 \ldots d\beta_p d\mu = 1,$$

kde $K \subset \boldsymbol{R}^{p+2}$ je množina hodnot, kterých mohou nabývat jednotlivé parametry. Přesný výpočet je však výpočetně náročný a jeho náročnost prudce roste s rostoucím počtem odhadovaných parametrů. Zpravidla se proto využívají simulační metody Monte Carlo.

Výsledkem bude tzv. *posteriorní rozdělení* parametrů, které zpřesní, případně upraví naše apriorní přesvědčení o rozdělení pravděpodobnosti těchto parametrů. Můžeme tak například zkoumat, s jakou pravděpodobností má nějaký regresor pozitivní vliv na přežívání:

$$P(\beta_j < 0) = \int_J \int_{-\infty}^{0} f_{posterior}(\alpha, \boldsymbol{\beta}, \mu) d\beta_j d\alpha d\beta_1 \dots d\beta_{j-1} d\beta_{j+1} \dots d\beta_p d\mu,$$

kde $J \subset \boldsymbol{R}^{p+1}$ značí množinu hodnot, kterých mohou nabývat ostatní parametry.

Nás bude zajímat především hazardní funkce, k jejímuž nalezení ale nejprve musíme znát funkci přežití. Tu dostaneme jako:

$$S(t) = 1 - F(t) = 1 - \int_0^t \frac{\alpha}{\lambda}\Big(\frac{x}{\lambda}\Big)^{\alpha-1} e^{-(\frac{x}{\lambda})^{\alpha}} dx.$$

Uvedený integrál spočítáme pomocí substituce $s = -\big(\frac{x}{\lambda}\big)^{\alpha}$, $ds = -\frac{\alpha x^{\alpha-1}}{\lambda^{\alpha}}$. Funkce přežívání se zjednodušší do tvaru

$$S(t) = 1 - \int_0^{-(\frac{t}{\lambda})^{\alpha}} e^s = 1 - \big[e^s\big]_0^{-(\frac{t}{\lambda})^{\alpha}} = e^{-(\frac{t}{\lambda})^{\alpha}}.$$

Nyní už můžeme spočítat hazardní funkci jako

$$h(t, x, \boldsymbol{\beta}, \mu, \alpha) = -\frac{\partial \ln(S_t)}{\partial t} = -\frac{\partial - (\frac{t}{\lambda})^{\alpha}}{\partial t} = \lambda^{-\alpha}\alpha t^{\alpha-1}, \tag{2.16}$$

kde za λ dosadíme ze vztahu 2.14 a dostaneme

$$h(t, x, \boldsymbol{\beta}, \mu, \alpha) = \alpha t^{\alpha-1}\big(e^{-\frac{\mu+\boldsymbol{x'\beta}}{\alpha}}\big)^{-\alpha} = \alpha t^{\alpha-1}\big(e^{\frac{\mu}{\alpha}}\big)^{\alpha} e^{\boldsymbol{x'\beta}} \tag{2.17}$$

Zavedeme-li substituci $\gamma = e^{\frac{\mu}{\alpha}}$, dostaneme častější tvar zápisu hazardní funkce pro Weibullův model přežívání:

$$h(t, x, \boldsymbol{\beta}, \gamma, \alpha) = \alpha\gamma(t\gamma)^{\alpha-1} e^{\boldsymbol{x'\beta}}.$$

Kapitola 3

Analýza přežívání na datech o nádorových onemocněních v dutině ústní

3.1 Klasická analýza přežívání

Nejprve se pojďme podívat na rozdělení pacientů do rizikových skupin. Na základě výsledků publikovaných v různých odborných článcích jsme pacienty rozdělili čtyřmi různými způsoby do tří rizikových skupin a nás bude nyní zajímat, které dělení nejlépe odpovídalo úmrtnosti pacientů v našich datech. K posouzení, který z modelů je lepší, můžeme použít například *Akaikeho informační kritérium*, které zohledňuje věrohodnost modelu a také počet parametrů v tomto modelu. Obecně se spočítá jako

$$AIC = -2\ln(L) + 2p,$$

kde $\ln(L)$ je logaritmus maxima funkce věrohodnosti a p značí počet parametrů modelu.

Jednotlivé proměnné *risk group* nabývají tří hodnot kódovaných 1, 2 a 3, což můžeme pokládat za rozdělení na slabě ohrožené, středně ohrožené a silně ohrožené jedince. Při sestavování tohoto modelu pak můžeme postupovat dvěma způsoby. Jestliže budeme předpokládat, že rozdíl mezi slabě a středně ohroženými jedinci je stejný jako mezi středně a silně ohroženými, můžeme tuto proměnnou uvažovat jako numerickou a dostaneme tak jednoduchý model s jediným re-

	Risk group jako faktor	Risk group jako číslo
Dělení 1	1041.4	1039.7
Dělení 2	1025.1	1024.9
Dělení 3	1033.7	1035.3
Dělení 4	1036.5	1034.7

Tabulka 3.1: Tabulka hodnot Akaikeho informačního kritéria pro jednotlivá dělení do rizikových skupin

gresorem, který bude nabývat těchto tří hodnot. My bychom ale spíše s touto proměnnou chtěli pracovat bez této omezující podmínky a budeme ji tedy vnímat jen jako kategoriální. V takovém případě musíme do modelu zavést dvě umělé proměnné (o jednu méně, než kolika hodnot může regresor nabývat). Coxův model pak bude vypadat následovně:

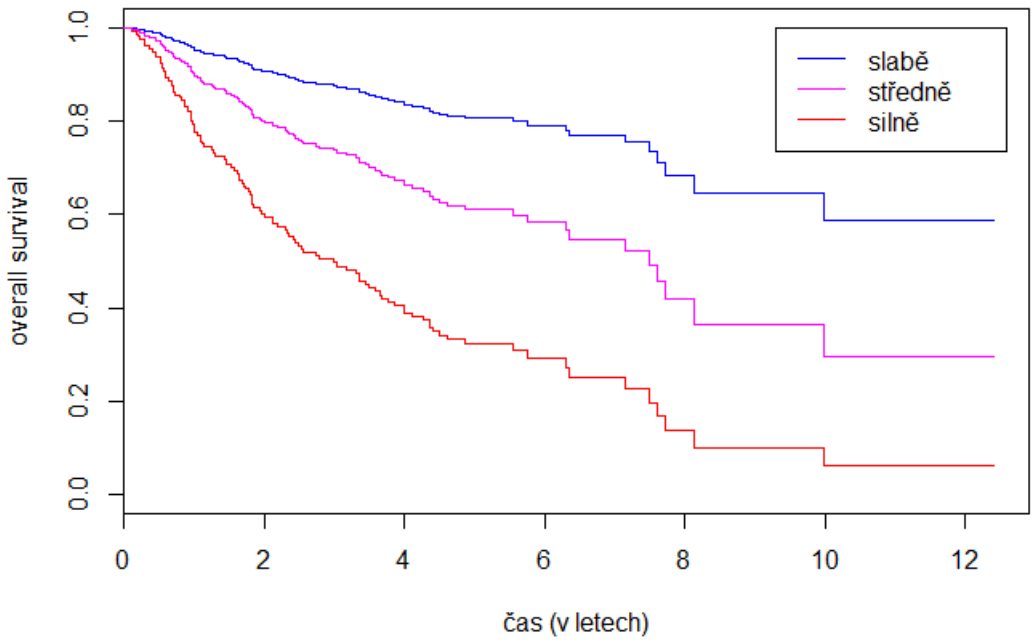
$$h(t|RG) = h_0(t)e^{\beta_1 I_{[RG=2]}+\beta_2 I_{[RG=3]}},$$

kde RG značí regresor *risk group* a I tzv. indikátorovou funkci, která je rovna jedné, jestliže je podmínka v argumentu splněna, a nule, jestliže ne. Pozorování, která spadají do rizikové skupiny *slabě ohrožení*, tvoří tzv. referenční skupinu, jelikož jejich hazardní funkci dostaneme, budou-li hodnoty u všech regresních parametrů β_i rovny nule. Hodnoty e^{β_1} resp. e^{β_2} nám budou vyjadřovat, kolikrát větší je hodnota hazardní funkce pro středně, resp. silně ohrožené jedince právě oproti této referenční kategorii. Budeme-li chtít znát poměr rizik také mezi středně a silně ohroženými, můžeme spočítat jejich hazard ratio tak, že do funkce e^x dosadíme rozdíl odhadů β_2 a β_1. Při konstrukci konfidenčního intervalu je pak potřeba mít na paměti, že

$$var(\hat{\beta}_2 - \hat{\beta}_1) = var(\hat{\beta}_2) + var(\hat{\beta}_1) - 2cov(\hat{\beta}_2, \hat{\beta}_1).$$

Oba uvedené modely sestavíme jak pro regresor s nominálním, tak pro regresor s ordinálním charakterem pro všechna čtyři rozdělení do rizikových skupin. Výsledky shrneme do tabulky 3.1

Nejnižší hodnoty AIC dosáhneme v obou modelech pro druhé dělení, které je tedy zřejmě nejvýstižnější, alespoň pro naše data. O něco lépe navíc vychází

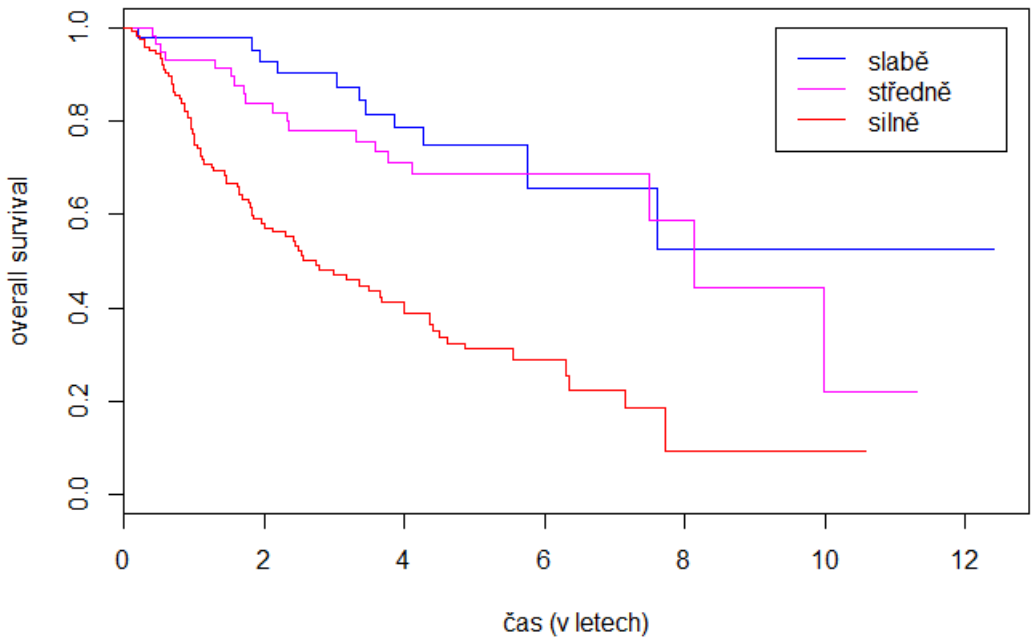


Obrázek 3.1: Srovnání Coxovy funkce přežití pro pacienty v jednotlivých rizikových skupinách

model, kde jsme ušetřili jeden parametr tím, že jsme uvažovali stejný rozdíl mezi slabě a středně ohroženými, jako mezi středně a silně ohroženými. Rozdíly, které jsou mezi první a druhou rizikovou skupinou a mezi druhou a třetí rizikovou skupinou, se tedy očividně výrazněji neliší. V tomto případě bychom poměr rizik mezi dvěma sousedními skupinami odhadli jako 2.29. Pacienti v rizikovější skupině jsou na tom tedy průměrně dvakrát hůř, co se týká doby přežití, oproti pacientům, kteří jsou o jednu rizikovou skupinu níž.

Odhadnutou Coxovu funkci přežití pro pacienty v jednotlivých rizikových skupinách (na základě druhého dělení) můžeme vidět v grafu na obrázku 3.1.

Vykreslit si můžeme také neparametrický Kaplan-Meierův odhad (na obrázku 3.2), který se bude do jisté míry lišit, neboť nemá předpoklady proporcionality hazardní funkce v každém časovém okamžiku. Celkový obraz o zkušenosti s přežíváním by se však neměl lišit nijak výrazně.

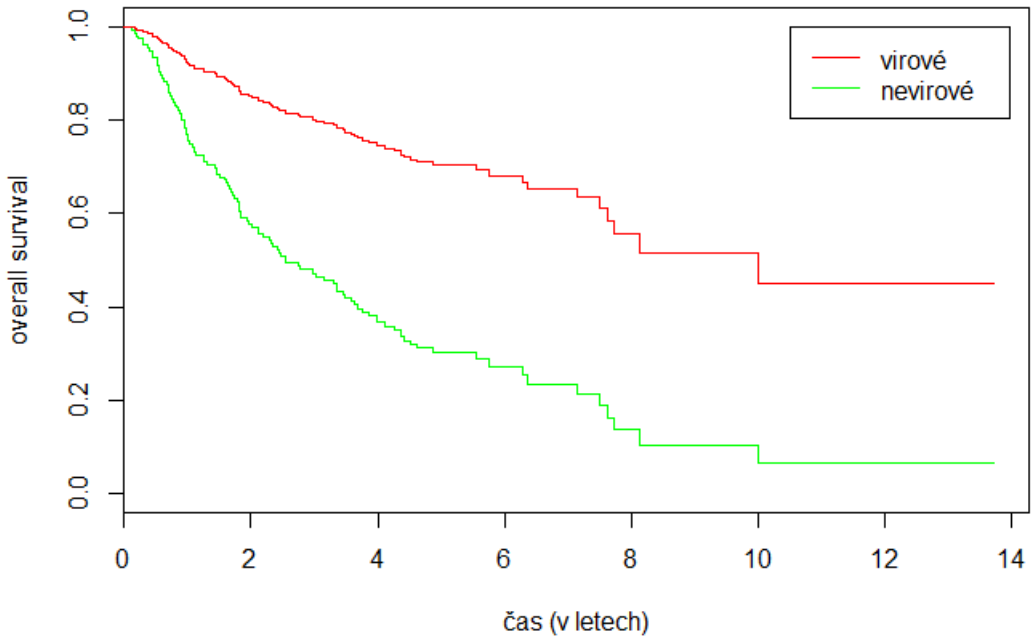


Obrázek 3.2: Srovnání Kaplan-Meierovy funkce přežití pro pacienty v jednotlivých rizikových skupinách

Zde můžeme vidět, že předpoklad proporcionality modelu může být někdy velmi přísný. Zatímco v Coxově modelu bychom na základě grafu tvrdili, že se zkušenost s přežíváním snižuje pro rizikovější skupiny, v případě Kaplan-Meierova odhadu bychom si zřejmě nebyli tolik jistí odlišností přežívání u slabě a středně ohrožených jedinců.

Nyní už přejdeme k sestavení vlastního modelu. Jednou z nejdůležitějších vysvětlujících proměnných v našem modelu bude informace o tom, zda je nádor u daného pacienta virového nebo nevirového původu, v další analýze se tedy podíváme blíže na to, jak se přežívání v těchto dvou skupinách liší. To, že přežívání mezi nimi bude rozdílné, můžeme vidět už z obrázku 2.1, kde jsme srovnávali kumulativní hazardní funkci pro obě skupiny a viděli jsme, že pro pacienty s nevirovým onemocněním toto riziko roste rychleji. Stejnou odpověď nám dá také odhad parametru β v Coxově modelu, kde prozatím nebudeme uvažovat

žádný další regresor. Po dosazení odhadu $\hat{\beta}$ do funkce e^x dostaneme hodnotu 0.294, která nám říká, že riziko úmrtí je pro pacienty s virovým onemocněním v kterémkoliv okamžiku přibližně třikrát méně než pro pacienty s nádorem nevirového původu. Odpovídající funkce přežití pro obě skupiny jsou vykresleny na obrázku 3.3.



Obrázek 3.3: Srovnání funkce přežití podle Coxova modelu pro pacienty s virovým a nevirovým nádorovým onemocněním

Hlavní otázkou, která nás bude zajímat, je, jestli má význam stanovit terapii na základě toho, jestli je nádor virového nebo nevirového původu. Obě skupiny jsou v datech zastoupeny v podobném počtu pacientů (114 pacientů má onemocnění virového původu, 119 nevirového) a s ohledem na to, že máme pro obě skupiny dostatek pozorování, můžeme si dovolit prozkoumat každou z nich zvlášť. Sestavíme tedy dva modely, které budou dosahovat nejnižšího Akaikeho informačního kritéria na příslušných datech. Výsledky shrneme do tabulky 3.2.

Virová onemocnění		Nevirová onemocnění	
Regresor	***HR***	**Regresor**	***HR***
Uzliny	1.602	**Uzliny**	1.292
Velikost	1.340	**Velikost**	1.379
Alkohol (užívá)	2.190	**Stádium**	0.550
Alkohol (užíval)	5.651	**Vzdělání**	0.9256
Věk v čase diagnózy	1.041		

Tabulka 3.2: Přehled odhadů poměrů rizik v Coxově modelu při rozlišení pacientů na pacienty s virovým a nevirovým nádorovým onemocněním

Ani v jednom případě bychom terapii neoznačili jako významný regresor (její odebrání z modelu snížilo hodnotu Akaikeho informačního kritéria). Stanovovat léčbu na základě toho, zda se jedná o virové nebo nevirové onemocnění, tedy nemá význam. Doba přežití zřejmě na zvolené terapii nezávisí. Co by nás však mohlo zaujmout, je hodnota parametru u regresoru *užíval alkohol v minulosti* u pacientů s virovým onemocněním. Ta nám totiž říká, že jestliže dotyčný užíval alkohol, ale přestal, je jeho šance na přežití pětkrát nižší než u člověka, který alkohol nikdy neužíval. To by samo o sobě zvláštní nebylo, ale pacienti, kteří alkohol nadále užívají, podle dat umírají více než dvakrát „pomaleji“. Zde patrně narážíme na obrovskou variabilitu v možnostech pití alkoholu. Uvedená hodnota parametrů je zřejmě důsledkem toho, že spousta lidí užívá alkohol v tom smyslu, že občas trochu alkoholu vypijí. Užíval alkohol, ale přestal, se pak může týkat především lidí, kteří alkohol užívali ve velkém množství a často přestali právě kvůli zničenému zdraví. U onemocnění nevirového původu se nám nepodařilo prokázat, že by užívání alkoholu mělo významný vliv na přežívání.

Pro porovnání se ještě můžeme podívat, jak by vypadal nejlepší Coxův model (podle Akaikeho informačního kritéria) v případě, že bychom pracovali s celým datasetem a informaci o tom, zda je nádor virového či nevirového původu, bychom zahrnuli jen jako regresor. V takovém případě z modelu vystoupí vysvětlující proměnné *věk v čase diagnózy, uzliny, stádium* a *vzdělání*. Přehled významných regresorů je uveden v tabulce 3.3.

Regresor	HR
Velikost	1.257
Alkohol (užívá)	1.857
Alkohol (užíval)	3.487
Virové onemocnění	0.307

Tabulka 3.3: Přehled odhadů poměrů rizik v Coxově modelu aplikovaném na všechna analyzovaná data

3.2 Bayesovská analýza přežívání

Pro bayesovskou analýzu přežívání jsme využili platformu Stan[2], což je moderní nástroj pro statistické modelování a výpočty, který lze implementovat také do softwaru R. Ke své práci využívá Weibullovo rozdělení dané předpisem 2.13, kde předpokládáme, jak bylo uvedeno výše, že na regresních koeficientech závisí parametr λ.

Výsledky bayesovské analýzy budeme srovnávat s výsledky klasického Weibullova modelu, jehož parametry můžeme v softwaru R odhadnout pomocí funkce *survreg()* v knihovně *survival*. Podle toho, jaké rozdělení pravděpodobnosti časů budeme předpokládat, můžeme zadat například parametr *dist="exponential"* nebo *dist="weibull"*. My budeme počítat s druhým případem. Je však potřeba mít na paměti, že funkce pro klasický model pracuje s nepatrně odlišnou podobou závislosti mezi λ a regresními koeficienty[4], totiž

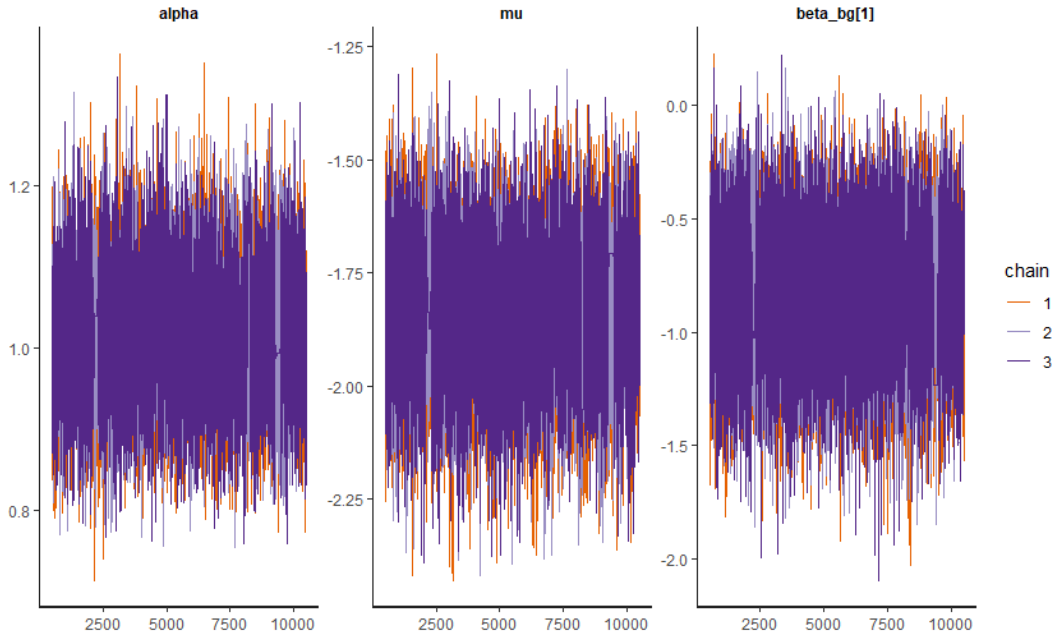
$$\lambda_i = \exp(\mu + \boldsymbol{x}'\boldsymbol{\beta}),$$

a tedy koeficienty, které nám software poskytne, jsou $-\boldsymbol{\beta}\alpha$ (to lze snadno odvodit, dosadíme-li uvedený odhad λ_i do předpisu hazardní funkce 2.16). K hodnotám koeficientů $\boldsymbol{\beta}$, tak abychom mohli $e^{\boldsymbol{\beta}}$ interpretovat jako poměry rizik (a zároveň abychom dostali odhad stejného parametru, jehož posteriorní rozdělení pravděpodobnosti se snažíme najít v bayesovském modelu), se proto dostaneme, když odhady koeficientů ze softwaru vydělíme odhadem parametru α a vynásobíme mínus jednou.

Pro jednoduchost začneme analýzou modelu s jedinou proměnnou nabývající pouze dvou hodnot, např. budeme uvažovat jen regresor pohlaví, kde 0 bude

značit muže a 1 ženy. V případě klasického Weibullova modelu dostaneme, že odhad příslušného parametru $-\alpha\beta$ je přibližně 0.813 a parametr α bychom odhadli jako 0.985. Tedy regresní parametr β odhadneme jako $\hat{\beta} = -\frac{0.813}{0.985} = -0.801$.

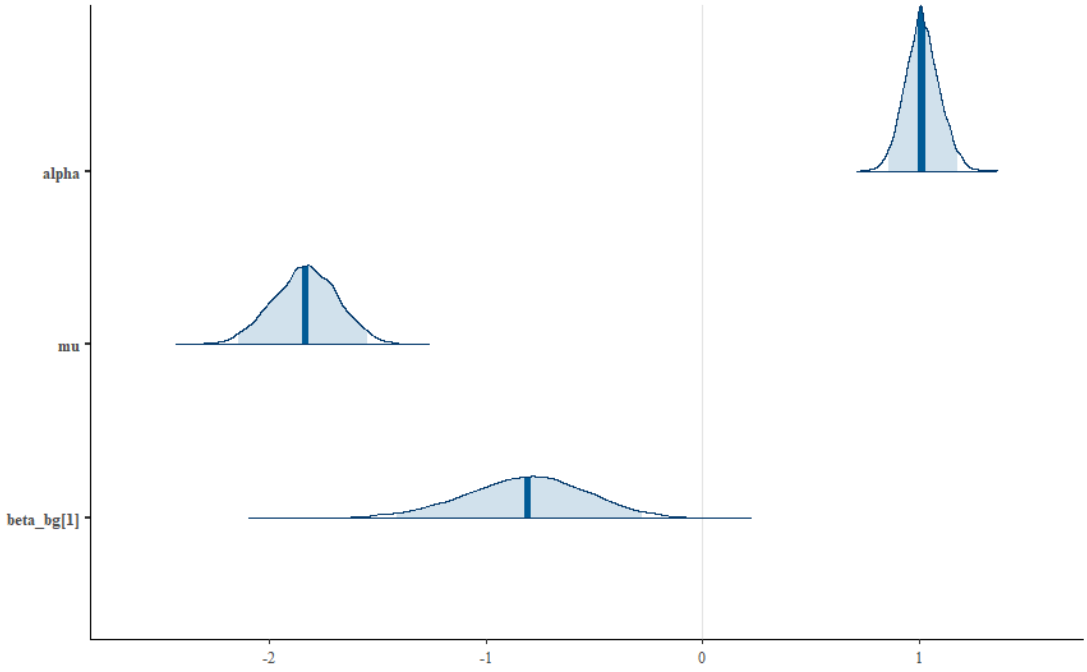
Nyní přejdeme k bayesovské analýze. Vytvoříme si v platformě Stan Weibullův model pro analýzu přežívání, kde bude jediný regresor pohlaví, a nasimulujeme hodnoty parametrů v tomto modelu. Z pomocí funkce *traceplot()* v balíčku *rstan* si můžeme vykreslit průběh simulace do grafu na obrázku 3.4.



Obrázek 3.4: Průběh simulace hodnot z posteriorních rozdělení pravděpodobnosti parametrů z Weibullova modelu s jediným regresorem pohlaví

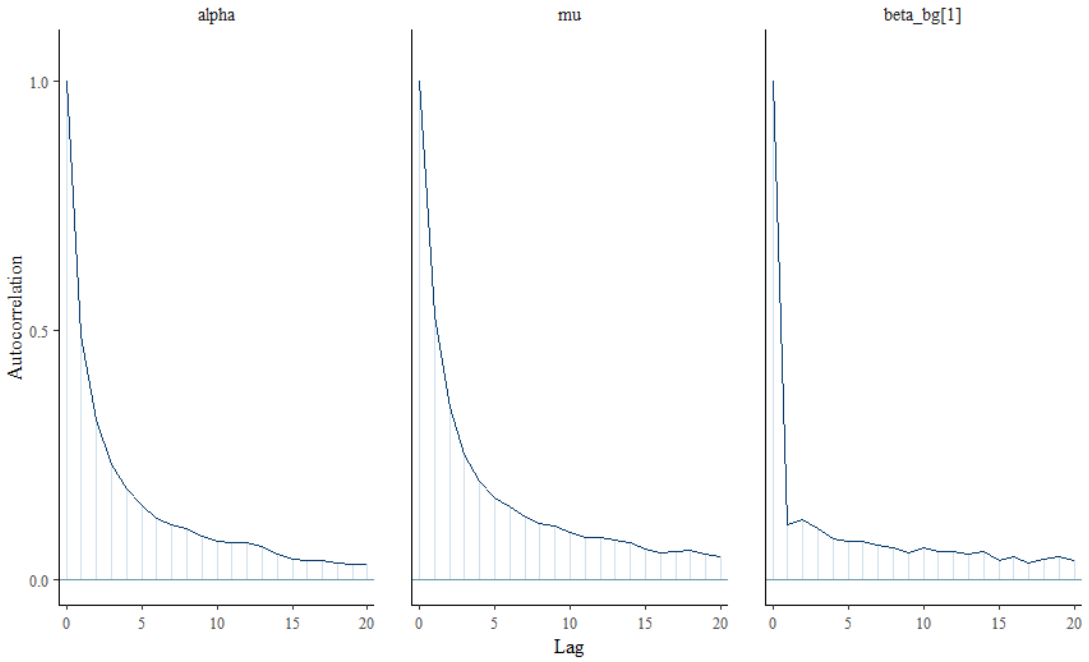
V uvedeném grafu jsou vykreslené tři simulace s různými výchozími body, abychom tak měli větší jistotu, že počáteční volba neovlivní výsledek. Z toho, že se všechny tři řetězce simulací nacházejí z velké části v zákrytu, můžeme usuzovat, že volba výchozích hodnot pro simulaci výsledek příliš neovlivnila, a současně také můžeme vidět, že skutečná hodnota parametru β, který nás zajímá především, se s největší pravděpodobností nachází někde mezi -1.5 a -0.25, tedy pohlaví zřejmě má vliv na přežívání, a to v tom smyslu, že ženy mají riziko úmrtí

podstatně nižší než muži a toto riziko je pro ně v kterémkoliv okamžiku 0.2 až 0.8-násobkem rizika u muže (k uvedeným hodnotám se dostaneme, dosadíme-li oba odhady krajních bodů do funkce e^x). O něco přesnější představu získáme, když si vykreslíme odhadnutá posteriorní rozdělení parametrů s pomocí funkce *mcmc_areas* z balíčku *bayesplot*.



Obrázek 3.5: Posteriorní rozdělení pravděpodobnosti parametrů z Weibullova modelu s jediným regresorem pohlaví

Ještě než se pustíme do hlubší analýzy výsledků, podotkněme, že u simulování hodnot z pravděpodobnostního rozdělení nedostaneme přímo nezávislé hodnoty, ale my bychom chtěli, aby byly generované hodnoty co nejméně korelované. Představu o tom, jak dobře nebo špatně na tom jsme, můžeme získat s pomocí funkce *mcmc_acf*, která je rovněž v balíčku *bayesplot*, a která nám vykreslí graf pro autokorelaci jednotlivých hodnot. Pokud by byla autokorelace vyšších řádů příliš vysoká, měli bychom simulaci zopakovat s jinou volbou počátečního bodu. Graf autokorelace vidíme na obrázku 3.6 a zřejmě bychom tedy výsledky simulace mohli považovat za věrohodné.



Obrázek 3.6: Autokorelační funkce při simulování hodnot z posteriorních rozdělení parametrů

Pokud bychom chtěli jen nějakou obecnou představu o hodnotě parametru β, mohli bychom spočítat průměr z jeho nasimulovaných hodnot, a získat tak odhad střední hodnoty tohoto parametru. Výsledný odhad je -0.817, což je hodnota velmi blízká k odhadu pomocí klasického modelu (-0.801).

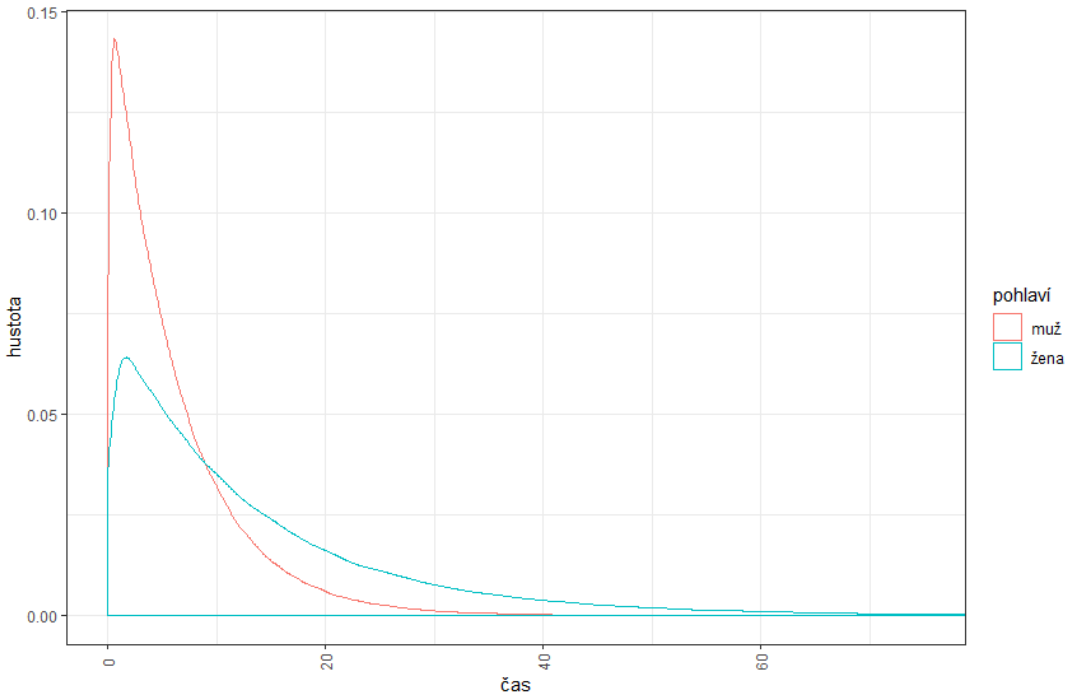
Vzhledem k tomu, že máme odhad celého posteriorního rozdělení parametru β, máme ale možnost ptát se na odhad pravděpodobnosti řady jevů. První otázkou, která se nabízí, je, jaká je pravděpodobnost, že je parametr β opravdu menší než nula, tedy že ženy přežívají v průměru déle než muži. V takovém případě bychom dostali dost přesvědčivou odpověď, totiž že tato pravděpodobnost je přibližně 0.9992. S ohledem na to, že chceme srovnávat klasický a bayesovský model, hodí se spíše otázky typu, jaká je pravděpodobnost, že je rozdíl mezi přežíváním u mužů a u žen ještě větší, než jak jsme ho odhadli podle klasického modelu (v tomto případě tedy jaká je pravděpodobnost, že je parametr β ještě menší než -0.801). Odpovídající pravděpodobnost je 0.474, což koresponduje s naším očekáváním, že nám klasický model poskytne přibližně odhad střední

Chyba v % z $\hat{\beta}$	$P\big(\beta \in (\hat{\beta} \pm chyba)\big)$
5 %	0.115
10 %	0.229
20 %	0.441
25 %	0.531
40 %	0.751
50 %	0.853
60 %	0.917

Tabulka 3.4: Přehled pravděpodobností toho, že klasický model neodhadl parametr hůř než s danou chybou

hodnoty parametru β. Zřejmě nejzajímavější otázkou však bude, jaká je pravděpodobnost, že se skutečná hodnota parametru neliší od odhadu z klasického modelu o více než o nějakou danou chybu, např. o 10 %. V uvedeném případě bychom tedy odhadovali pravděpodobnost, že skutečná hodnota parametru leží v intervalu od -0.8811 do -0.7209 (hodnoty jsme získali přičtením a odečtením 10% z -0.801 k odhadu parametru z klasického modelu). S tím bychom zřejmě příliš spokojení nebyli, poněvadž tato pravděpodobnost je pouze 0.229. Uberemeli z přísnosti a dovolíme klasickému modelu chybu až do výše 25 % odhadnuté hodnoty, odhadneme pravděpodobnost, že se skutečný parametr nachází v tomto intervalu, už na 0.531. Přehled dalších pravděpodobností, s jakými se skutečná hodnota parametru nachází v intervalu *klasický odhad*$\pm$*chyba*, je v tabulce 3.4.

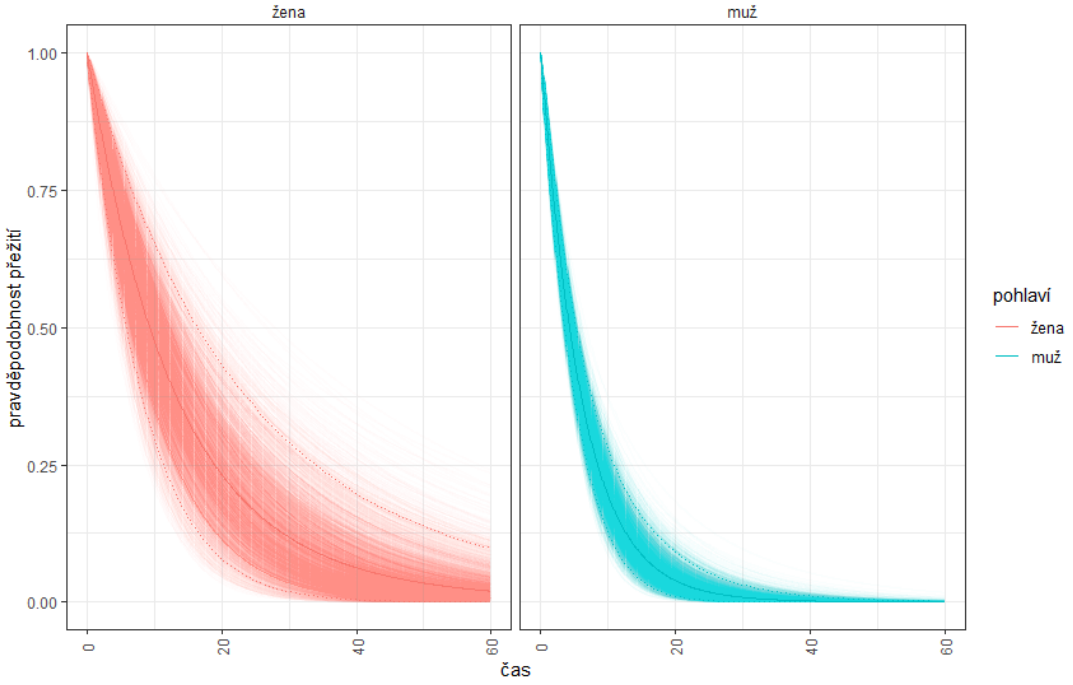
Kromě parametru β nám bayesovský model umožňuje také zabývat se pravděpodobností, s jakou pacient přežije po určitou dobu, a vykreslit si funkci hustoty doby přežívání pro muže i pro ženy (obrázek 3.7).



Obrázek 3.7: Odhad hustoty doby přežívání pro muže (červeně) a pro ženy (modře)

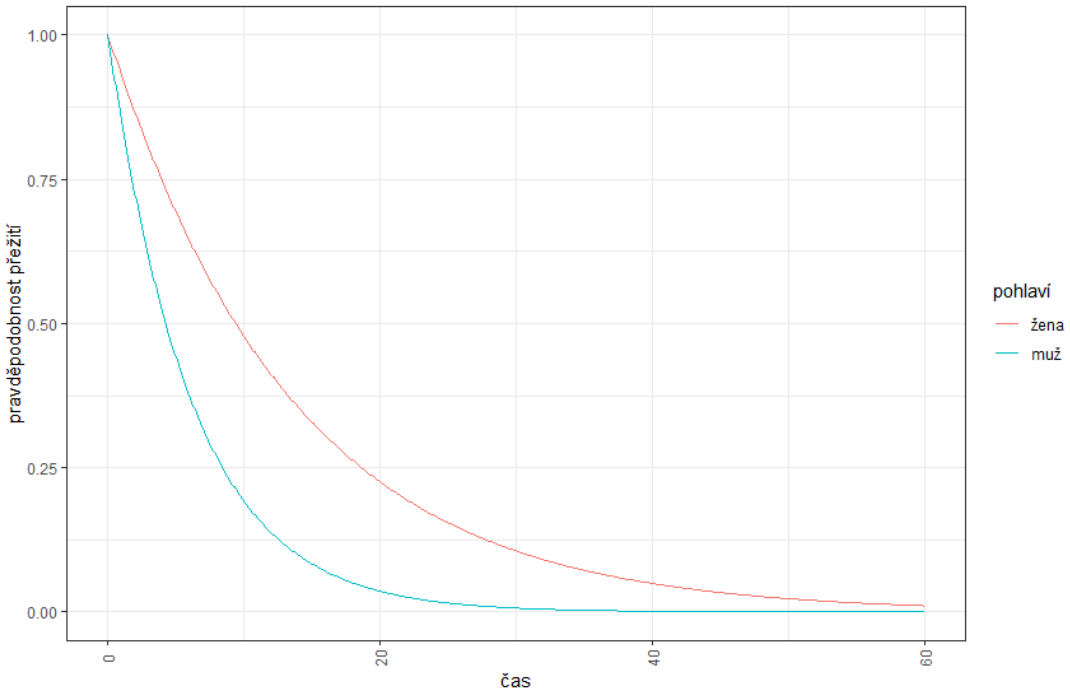
Z grafu můžeme opět vidět, že ženy jsou na tom lépe, co se týká přežívání, tedy že riziko úmrtí je pro ně podstatně nižší než pro muže. Střední dobu přežití bychom odhadli pro muže přibližně jako 6 let, pro ženy dokonce 14 let. Spočítáme-li také medián, zjistíme, že ženy mají přibližně 50% možnost přežití po dobu devíti let, muži se stejnou pravděpodobností přežijí jen 4 roky. Zřejmě mnohem užitečnější informaci nám poskytne odhad 10% kvantilu, který nám řekne, kolik let nejméně pacient přežije s 90% pravděpodobností. Pro ženy odhadneme příslušný kvantil jako 1.4, pro muže jako 0.6 let.

Kromě hustoty doby přežívání můžeme vykreslit také posteriorní odhad funkce přežívání, k čemuž využijeme v softwaru R funkci *ggplot*, která je součástí knihovny *ggplot2*. Odhad funkce přežití vykreslíme zvlášť pro ženy a zvlášť pro muže.



Obrázek 3.8: Posteriorní odhad funkce přežívání pro ženy (červeně) a pro muže (modře)

Z grafu na obrázku 3.8 můžeme opět pozorovat, že muži jsou na tom s dobou přežívání hůře než ženy, ačkoliv u žen můžeme vidět mnohem větší nejistotu v odhadu funkce přežívání, tedy mnohem širší rozpětí 95% věrohodnostního intervalu než u mužů. Pro zjednodušení a lepší ilustraci rozdílu v přežívání pro obě pohlaví můžeme vykreslit funkci přežívání pro obě pohlaví do jednoho grafu s tím, že parametry zvolíme jako odhady střední hodnoty jejich posteriorního rozdělení pravděpodobnosti. Výsledek je vidět na obrázku 3.9.



Obrázek 3.9: Funkce přežívání pro muže (modrá) a pro ženy (červená) při odhadech parametrů pomocí odhadů středních hodnot jejich posteriorního rozdělení pravděpodobnosti

Závěr

V diplomové práci jsme nejprve představili data, s nimiž jsme dále pracovali, a poté uvedli i několik základních pojmů z analýzy přežívání – vysvětlili jsme si, že pod událostí budeme v našem případě rozumět úmrtí, ačkoliv ne vždy to tak musí být, a také, že rozlišujeme několik druhů úmrtí, například úmrtí z jakékoliv příčiny a úmrtí z příčiny dané nemoci, přičemž použití obou definic je ospravedlnitelné. V další práci jsme uvažovali za událost úmrtí z jakékoliv příčiny. Poté jsme si objasnili pojem cenzorování a vysvětlili si nezbytnost zahrnutí cenzorování do výpočtu. V závěru první kapitoly jsme pak ještě vysvětlili pojmy funkce přežívání, která vyjadřuje pravděpodobnost přežití jedince různě dlouhá časová období, a hazardní funkce, která naopak vypovídá o pravděpodobnosti úmrtí v daném časovém okamžiku, jestliže se pacient dožil okamžiku těsně před.

Ve druhé kapitole jsme si nejprve představili klasickou analýzu přežívání, především pak nejčastěji používaný Coxův model, nastínili jsme však i podobu exponenciálního a Weibullova modelu. Coxovu modelu jsme se věnovali více a uvedli jsme u něj také způsob odhadu regresních parametrů β, jejich interpretaci s pomocí poměru rizik HR, testy významnosti i konstrukci konfidenčních intervalů. Kapitolu jsme začínali s případem s jedinou vysvětlující proměnou, který jsme poté zobecnili na případ s více regresory, pro nějž jsme opět uvedli také jeden z možných testů nulovosti celého vektoru regresorů, tj. testu nulové hypotézy $H_0 : \beta_1 = \cdots = \beta_p = 0$. V této kapitole jsme si také uvedli, jak lze odhadnout celou funkci přežívání pro Coxův model.

Závěr druhé kapitoly jsme věnovali bayesovské analýze, pro níž jsme si představili základní myšlenku odhadu aposteriorního rozdělení. Navíc jsme si více

přiblížili Weibullův model a jeho propojení s Weibullovým rozdělením časů přežití, které jsme uvažovali pro bayesovskou analýzu, jelikož Weibullův model narozdíl od Coxova modelu spadá do parametrických modelů.

Ve třetí kapitole jsme podrobili data nejprve klasické analýze přežívání a našli jsme pomocí Akaikeho informačního kritéria nejlepší Coxův model zvlášť pro pacienty s virovým a nevirovým původem onemocnění. Hlavní otázkou, která nás zajímala, bylo jestli má na přežívání pacientů vliv použitá terapie, tento regresor jsme ovšem v obou případech vyřadili z modelu a tedy považujeme vliv terapie za statisticky nevýznamný. Jako významné regresory vyšly v případě virových onemocnění uzliny, velikost nádoru, užívání alkoholu a věk pacienta. Pro nevirová onemocnění to byly uzliny, velikost nádoru, stádium onemocnění a vzdělání.

Poté jsme data analyzovali také bayesovsky s využitím Weibullova modelu. Pro jednoduchost a kvůli výpočetní náročnosti jsme uvažovali jen jediný parametr, a to pohlaví. Ten jsme nejprve odhadli s pomocí klasické analýzy přežívání, abychom mohli výsledky srovnávat. Poté jsme odhadli posteriorní rozdělení parametru β s využitím bayesovské analýzy a spočítali odhad jeho střední hodnoty, který vycházel blízký odhadu s využitím klasické analýzy. Pomocí posteriorního rozdělení pravděpodobnosti jsme také byli schopni odhadnout, že pravděpodobnost, že by muži přežívali lépe než ženy, je menší než 0.001. Také jsme se podívali na odhad hustoty doby přežívání pro ženy a muže, z níž bylo opět patrné, že ženy mají větší šanci na přežití delší dobu. Zatímco pro muže vycházela střední doba přežití přibližně 6 let, pro ženy vycházela přibližně 14 let. Na závěr jsme se podívali také na odhad funkce přežívání, z níž bylo opět patrné, že ženy jsou na tom, co se týká přežívání, lépe. Byla ale také vidět znatelně větší nejistota v odhadu než u funkce přežívání pro muže.

Literatura

[1] Hosmer, D., Lemeshow, S.: *Applied Survival Analysis: regression modeling of time to event data.* New York: Wiley, 1999. ISBN 0-471-15410-5.

[2] Rpubs – Bayesian Survival Analysis 1: Weibull Model with Stan, https://rpubs.com/kaz_yos/bayes_surv1 [cit. 26. 4. 2019].

[3] Weibull distribution – Wikipedia, https://en.wikipedia.org/wiki/Weibull_distribution [cit. 27. 4. 2019].

[4] Parametric regression model for survival data: Weibull regression model as an example, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5233524/ [cit. 27. 4. 2019].

[5] LAWLESS, Jerald F.: *Statistical models and methods for lifetime data.* New York: Wiley, 1982. ISBN 0471085448.

[6] Vojtechova, Z., Sabol, I., Salakova, M., Turek, L., Grega, M., Smahelova, J., Vencalek, O., Lukesova, E., Klozar, J., Tachezy, R.: *Analysis of the integration of human papillomaviruses in head and neck tumours in relation to patients' prognosis.* International Journal of Cancer 2016, 138(2), 386-395.